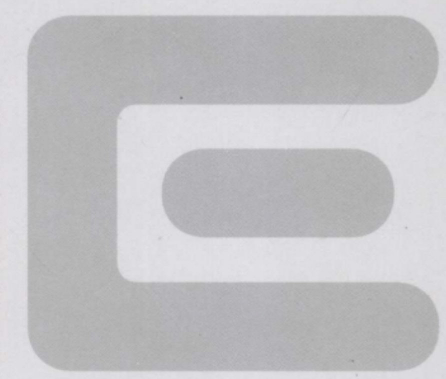
万卷方法 | 心理学研究方法丛书
中国心理学会心理学教学工作委员会推荐

DOING PSYCHOLOGY EXPERIMENTS



如何做心理学实验



大卫·W. 马丁(David W. Martin) 著
丁锦红 等译



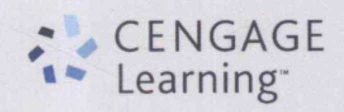


RUHE ZUO XINLIXUE SHIYAN

祝贺你随着我做心理学实验的思路学习到这里。希望我的语言和图示不仅不会妨碍你的进步，还能引起你的兴趣。信息量与精确度之间的平衡很是微妙——这个平衡对于每个读者是不同的。我希望我的行文不会让你太烦。

很显然，这本书不会把你变成一个现成的实验心理学家，但我相信，它一定提供了足以使你做一些简单实验的信息。你将发现，做实验比阅读怎样做实验有趣的多。所以，现在来做一些有趣的事吧！

——作 者

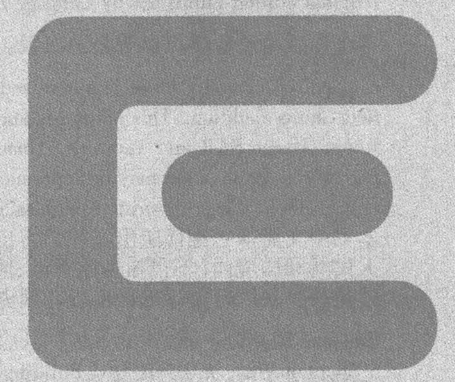


上架建议：学术社科
ISBN 978-7-5624-6151-7



定价: 45.00元

万卷方法 | 心理学研究方法丛书
中国心理学会心理学教学工作委员会推荐



如何做心理学实验

大卫·W. 马丁(David W. Martin) 著
丁锦红 等译

重庆大学出版社

Doing Psychology Experiments　978-0-495-11577-9
David W. Martin

Cengage Learning Asia Pte. Ltd.
5 Shenton Way, # 01-01 UIC Building, Singapore 068808

版贸核渝字(2008)第 098 号

图书在版编目(CIP)数据

如何做心理学实验/(美)马丁(Martin,D. W.)著;
丁锦红等译. —重庆:重庆大学出版社,2011.6
(万卷方法)
书名原文:Doing Psychology Experiments
ISBN 978-7-5624-6151-7

Ⅰ.①如…　Ⅱ.①马…②丁…　Ⅲ.①实验心理学
Ⅳ.①B84

中国版本图书馆 CIP 数据核字(2011)第 084585 号

如何做心理学实验
大卫·W.马丁(David W. Martin)　著
丁锦红　等译
策划编辑:雷少波
责任编辑:李桂英　　版式设计:雷少波
责任校对:夏　宇　　责任印制:赵　晟
*
重庆大学出版社出版发行
出版人:邓晓益
社址:重庆市沙坪坝正街 174 号重庆大学(A 区)内
邮编:400030
电话:(023) 65102378　65105781
传真:(023) 65103686　65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn (营销中心)
全国新华书店经销
重庆升光电力印务有限公司印刷
*
开本:787×1092　1/16　印张:18.25　字数:409千
2011 年 6 月第 1 版　　2011 年 6 月第 1 次印刷
印数:1—4 000
ISBN 978-7-5624-6151-7　定价:45.00 元

万卷方法学术委员会

作译者简介

大卫·W. 马丁(**David W. Martin**) 北卡罗莱纳大学心理系教授、系主任,并曾经任新墨西哥州立大学心理系教授和系主任。他在 Hanover College 获得心理学和物理学学士学位,在俄亥俄州立大学分别获得工程心理学硕士和博士学位。他讲授的课程包括实验方法、心理学引论、人类绩效以及注意。他曾经在北卡罗莱纳大学和新墨西哥州立大学获得教学方面的奖励。

马丁博士在注意、决策和记忆方面发表了很多学术论文。他是美国心理学会、美国心理协会、人的因素及工效学会以及心理测量学会的会员。他曾经是 Rocky Mountain 心理联合会的主席。

在闲暇时间,Martin 博士也乐得其所,他喜欢潜水、吹号、唱歌以及与两个儿子玩耍。在过去的 12 年中,他经常参加家用车越野比赛,被誉为"危险的 David,比赛教授"。

丁锦红 男,教授、博士生导师。1989 和 1992 年在杭州大学心理系获学士和硕士学位,1998 年在中国科学院心理研究所获博士学位;现任首都师范大学教育科学学院副院长兼心理系主任,中国心理学会教学工作委员会委员,中国心理学会科普工作委员会副主任。发表论文 50 余篇。主要研究方向包括视觉信息加工和人类记忆过程。

总　序

自古以来,人类一直在探索着自己的内心世界:我到底是个什么样的人?为什么在许多方面我与周围人们如此相似,而在其他方面又如此不同?我们是怎样认知世界的?为什么有时记忆会错误有时记忆会很牢固?喜怒哀乐、爱恋、责任感是如何产生的?情绪能自我调控吗?意识是怎么回事?梦是怎么回事,它能预测未来吗?为什么一个人独处时和在群体中的行为是不一样的?我们是怎样理解语言的,又是怎样组织和表达语言的?如此等等的问题,从传说、神话、甲骨文中我们都能发现,可见,人类长期以来一直对理解自身的内心世界有着浓厚的兴趣。然而直至1879年冯特(Wilhelm Wundt, 1832—1920)在莱比锡大学建立了心理实验室之后,人类才开始借助科学方法来寻求这些问题的答案。通过精密严格的数据收集与事实分析来研究心理与行为,积累知识,从而发展出今日的心理学。而随着心理学的发展,其研究方法也发展起来。

心理学家的科学研究是一种自觉的、有目的地探索精神世界的求知活动。这种探索,不仅有理论,还有与理论有关的观点及方法、仪器等。当代心理学主要有五种理论取向在探讨人类内心世界的奥秘,并对大脑如何工作提出了不同的假设,因而它们所采用的方法也不同。持生物学理论取向的心理学家认为心理是脑的机能,采用脑电图(EEG)、正电子发射断层扫描技术(PET)、功能核磁共振成像技术(FMRI)等方法来探讨人类是如何产生知觉、记忆、推理、情绪和某些人格特征的。持学习理论取向的心理学家认为心理是个体对环境条件作用积累经验的生理变化,采用操作条件作用、奖赏、惩罚、观察学习等方法来探讨人类和动物行为的形成及矫正。持认知理论取向的心理学家把人类的心理活动视为类似于计算机的信息加工,用反应时、正误率和口语报告等手段来探讨人类的知觉、记忆、言语和思维等心理过程。持精神分析论取向的心理学家把人的心理视为潜意识本能的表现,采用个案调查、诠释学方法来探讨个人内部的驱力、冲突或心理疾病等潜意识活动。持人本主义理论取向的心理学家把人的心理视为自我实现需求的表现,采用相关法、诠释学方法来探讨自由意志、个人成长、潜能实现等问题。在心理学研究中理论观点与方法始终是结合在一起,相辅相成的。它们既是指导这种探索活动的武器,又是保证这种活动取得成果的基础。正因为有了这一套套理论观点与方法的有机结合,心理学的科学研究才成为一种自觉的、有目的的定向活动,心理学也才成其为科学。因此,学习心理学研究方法具有十分重要的意义。

首先,有助于我们自觉地将理论与方法相结合,养成科学思维的习惯。心理学家的科学研究都具有明确的目的,即需要解决的问题,为此,就要对该问题以往的研究和目前的现状进行文献综述,并按照一定的有效程序对其进行探讨。即是说,心

理学研究的基本程序与任何科学研究是一样的，都包含下列步骤：选题和提出假设——设计研究方案（用以检验假设的真伪）——收集资料——整理和分析资料——解释结果和检验假设。从心理学研究程序的各个环节可以看出，在心理学研究中，精密的仪器和先进的实验设备固然重要，然而最重要的还是研究者的头脑。通过对心理学研究方法的学习，将有助于我们养成善思考和科学思维的习惯。心理学研究方法，不仅可以帮助我们运用自己的智慧去进行科学研究，而且它还可以帮助我们去鉴别自己和他人的研究成果的正确与谬误。正因为如此，有经验的学者在评价一篇学术论文时往往不只是看它的结论，而且，甚至是更重要的还要看论文作者是通过怎样的途径和方法而获得结论的。

其次，有助于激发我们的创新观念和达成创新目标。心理学的理论、观点和研究方法是多样性的，其研究成果的科学性也不同。学习了心理学研究方法之后，我们了解到心理学研究的各种方法，如个案法、相关法和实验法在心理学研究的不同时期有不同的用途，其信效度也是不同的。个案研究在心理学研究初期是有用的（有助于发现可供研究的现象和变量），但是要确定变量之间的因果关系，建立科学理论，则必须借助于实验法。弗洛伊德（Sigmund Freud, 1856—1939）根据自己对歇斯底里病人的临床观察和对梦、失误和笑话等的现象分析，建立起以潜意识动机为基础的精神分析理论。这个理论不是一个科学的理论，其中许多概念和命题缺乏实证效度。然而，熟悉实验法的学者想到用实验法来检验弗氏理论中的许多概念。例如，对于潜意识这种现象，他们发展出一系列的内隐实验程序来进行检验，结果发现有内隐学习、内隐记忆、内隐情绪、内隐动机的存在。虽然20世纪80年代兴起的心理潜意识研究与弗洛伊德的潜意识的性本能和死亡本能有本质的区别，但却加深了我们对心理成分和潜意识性质的认识。自我是心理学研究的一个主题，在科学心理学的早期，一些著作从理论上探讨自我的性质，到了20世纪70年代不少学者开始用相关法和实验法探讨自我的成分和机能，后来由于引入神经科学方法（如脑的电刺激、功能核磁共振成像技术）才开始探索自我的脑机制。正如巴甫洛夫（И. П. Павлов, 1849—1936）所说“科学是随着研究方法所获得的成就而前进的。研究方法每前进一步，我们就更提高一步。随之在我们面前也就开拓了一个充满着种种新的、更加广阔的远景。因此，我们头等重要的任务乃是制定研究方法。”①

第三，有助于年轻心理学工作者快速成长。做任何事情都要讲究方法。方法对头，事半功倍；方法不对，事倍功半，甚至导致失败。心理学史表明，有些心理学家之所以能在学科上有所建树、有所贡献，除了他们的天赋聪慧，当时的科技水平和良好的学术环境外，还往往与他们能正确运用新的研究方法有密切的关系。系统地学习心理学研究方法显然比只凭个人经验、漫无边际地去摸索，更能促进年轻心理学工作者的快速成长。

2005年秋天，重庆大学出版社“万卷方法”总策划雷少波同志带着已出版的新书来征求我对翻译这套图书的意见。我很高兴看到他们对这套书的设想：“万卷方法”是重庆大学出版社从2004年开始出版，拟系统深入地介绍各门社会科学研究方法的大型工具性丛书，其中包括心理学研究方法。这是一项促进我国心理学事业发

---

① 巴甫洛夫选集[M].北京：科学出版社，1955：49.

展的开创性工作，我给予热情鼓励和支持。在我看来，“万卷方法——心理学研究方法丛书”具有以下特点：

（1）品位高。对于研究方法的著作来说，质量优、品位高是最重要的。丛书所介绍的是国外心理学领域中，许多有成就的心理学家所普遍认可的心理学研究工作的原理、方法，研究课题设计，以及如何正确使用各种可以应用的技术手段等。例如如何做心理学实验、如何进行心理学的质性研究、如何撰写心理学学术论文，以及在心理研究中如何使用数理统计、应遵循哪些伦理道德等。其中像心理学质性研究、心理学研究伦理道德等著作，国内至今未见有专题著作翻译出版，是国内急需的。

（2）适用面广。丛书所介绍的心理学研究方法是相对基础性的，可供高校心理学专业的本科生和研究生作为教材或教学参考书使用，也可供广大人文社会科学工作者参考。

（3）开放性。根据我国心理学教学和科研的需要以及心理学研究方法的发展，出版社将通过版权引进和本土开发，使丛书不断丰富与完善。

我相信广大读者会喜爱“万卷方法——心理学研究方法丛书”，祝愿“万卷方法”不断发展，日益完善。

是为序。

黄希庭

谨识于2009年10月

西南大学 有容斋

译者前言

无论是开展心理学研究，还是将心理学付诸于实践，掌握心理学的研究方法是不可缺少的，尤其是心理学的实验方法。心理学实验方法并不抽象，但它必须具有严密的逻辑及必要的统计学知识；心理学实验方法也并不枯燥，但它需要你静下心来仔细分析与推敲。心理学的实验方法正是心理学成为科学的重要标志之一。因此，掌握心理学实验方法是从事心理学工作所必备的基本素质，它对解决实际问题也非常有益。在老师的指导下，系统掌握心理学实验方法或许相对容易。然而，一本好的教科书更是不可或缺，它可以更加发挥你的潜力和创造力。David W. Martin 教授的《如何做心理学实验》就是系统介绍心理学实验方法的教科书。它从初始的观察、形成实验构想、选择实验设计、实验的实施、实验数据分析以及实验报告的撰写等实验研究的所有方面进行了详细论述，为读者清晰地展现了一条通往实验研究之路。无论是心理专业的本科生还是研究生都能够非常容易地从该书中找到能够丰富自己心理学知识体系的必要构件。即使是初学者，也能通过该书构建起自己对心理学实验研究认识的基本框架。

在中国心理学会教学工作委员会及重庆大学出版社的推荐和组织下，我和我的同事们将 David W. Martin 教授的《如何做心理学实验》翻译成了中文版，以供国内心理学工作者及心理学专业学生参考与学习之用。译文难免有疏忽及欠妥之处，敬请各位同仁及同学指出，以促进我们共同成长。参加该书翻译的有（按章节顺序）：丁锦红（第 1 章）、王异芳（第 2 章）、刘肖岑（第 3 章）、汪亚珉（第 4、5 章）、王岩（第 6、7 章）、黄贤军（第 8、9 章）、陈婷婷（第 10 章）、魏萍（第 11、12 章）、卢珊（第 13 章、附录），其他内容由丁锦红翻译；全书由丁锦红审校。本人非常感谢参与本书翻译的诸位年轻博士们，他们在繁忙的工作中挤出时间，高质量地完成了译稿，他们的才华与合作精神打动了我，也激励着我。我还要感谢中国心理学会教学工作委员会同仁及重庆大学出版社雷少波、林佳木、李桂英编辑等的帮助与支持。

丁锦红

2011 年 1 月 31 日

作者前言

《如何做心理学实验》一书虽然出版至今已 30 年，但它在指导几乎或完全没有心理学基础知识的学生如何做简单的心理学实验方面似乎仍正发挥着独创性的作用。在这 7 个版本中，我试图保持着通俗易懂的写作风格。尽管应该采用客观的和非个人偏好的方式报告科学结果，但我相信，做实验是一种高度个人化的活动。实验者阅读文献并且形成对科学知识体系的观点。他们创立需要检验的理论和假设。他们选定需要控制和测量的变量。他们解释结果并决定如何完善科学知识体系。他们以个人的方式参与实验过程。因此，我认为，通过一本个人化的书籍是教会新手做实验的最好方式。

实际上，已有一些研究对学生使用更个性化课本的喜好作出了评价。例如，Paxton(1997)发现，学生在阅读一个"可视作者(visible author)"(以第一人称书写，介绍自己的个人经验)的作品时会与作者进行心灵交流，这将拉近读者与文字中信息之间的关系。密西根州立大学学生 Lorin Sheppard(2001)甚至研究了幽默在书中的作用，并将它与书中的其他内容进行了比较，她称之为"humorectomy"(个人交流资料，2001 年 3 月 27 日)。她发现，学生不仅觉得幽默章节更加有趣、信息丰富，而且在后来的回忆测验中也能够回忆出更多内容。我非常欣慰的是，这些结果支持我一直以来所持有的直觉，即幽默和个人写作风格在教学中是非常有用的。

现在，我简要介绍一下这本书的内容。它为那些没有实验背景知识和经验的学生学会设计、实施、解释以及报告简单的心理学实验提供了足够的知识。尽管这本书主要用于本科生的实验方法课程中，但它也可以与一些书一起在其他领域中使用。许多学校将它用在心理学导论课程的实验环节。它有时候也会与统计书一起被用在与统计、实验相关的课程中。在教师要求学生做实验而学生的实验背景知识非常缺乏的情况下，许多本科生的课程都使用本书。我曾经与很多读者(包括教师与学生)进行过交流。他们认为，这本书可以作为一本独立的课本，也可以作为补充读物。事实上，在我自己的实验方法课程中，我会在课程开始之前对学生进行一些小测验，测验题目来自于本书的题库，通过这种方法促使学生在课前阅读这本书。在课堂上，我重点讲解那些关键的内容，但更多的是讨论实验的计划和问题。本书成功地使原来层次不齐的学生最终达到相同的水平，而在课堂上可以开展更有创造性的互动交流。

尽管本书经常被当成补充课本，并且书的尺寸也比市场上其他类型的书要小，但它确实介绍了实验方法中多数重要概念。我也试图全面介绍本领域内的内容，一些研究表明，这种努力是成功的。①

心理学各个领域教科书的作者曾对自己领域中的术语、概念的重要性进行评估。在最重要的 100 个有关方法与统计的术语中，有 33 个是统计和心理测量方面

① Boneau, C. A. (1990). Psychological literacy: A first approximation. American Psychological Psychologist, 45, 891-900.

的术语，其余的67个是方法方面的术语。除了6个以外，本书中讨论了几乎所有这些与方法有关的术语。而这6个中的4个也在概念水平上用不同的说法进行讨论，只有2个未涉及。我相信，这些证据足以证明，本书涵盖了实验方法的主要内容。

本书没有涉及实验心理学领域中的具体内容和新进展。我使用的很多例子都是虚构的，它们只是用于描述所要讨论的内容，而不是真实的数据，因此，它们不会给学生一个有关实验心理学的全貌。但是，学生可以开始了解到他们在文献搜索中所涉及的内容，见第6章。本书也不会教给学生那些复杂的实验设计和统计分析方法。我尽量使得它简单明了。虽然我讨论了描述统计和推论统计的基本原理，但附录A中列出的实际统计操作也只是简明的菜谱而已。

本书的第7版有了很多变化。有些是微小的变化，如每章开头引用的名言、纠正了一些错误，以及增加了一些卡通画等。我改变了一些虚拟的例子，因为它们与真实的数据相矛盾，因此，有人提出反对意见。在第3章中，我增加了一个有关最新问题和终极问题之间差异的简短讨论。在第4章中，当我讨论使用“参与者(participant)”而不使用“被试(subject)”时，加入了一些反对的观点。我更新了动物伦理部分的内容。在第5章中，我对有关剽窃的讨论进行了扩展，增加了互联网剽窃和一些实例。同时，还增加了与研究有关的“美国心理学会心理学家道德原则和行为规范(*APA Ethical Principles of Psychologists and code of conduct*)”。在第6章中，我更新了电子资源搜索部分，增加了 *PsycINFO* 和 *PsycARTICLES*。在第12章中讨论推论统计逻辑时，我增加了虚无假设、Ⅰ类错误、Ⅱ类错误以及确定统计检验功效的方法等内容。在第13章中，我更新了在学术会议中做报告的内容，因为现在几乎所有报告都是在计算机上完成。许多人向我建议，应该提供一些有关在文章中如何报告统计结果的例子，因此，我在附录A中增加了一些这方面的例子。

此外，在第7版中扩充了为教师提供的考试题库。最新的网络资源见 http://www.thomsonedu.com/psychology/martin。

在这一版的修订中，我在保留必要内容的同时，尽量使本书保持短小精悍。事实上，此版本的页数比以往的版本少。我不希望这本书对学生而言过于富丽堂皇，而是希望它的价格能够适合于他们。对于那些用过本书以前版本的人而言，希望他们喜欢这些变化；对新读者而言，希望他们能够喜欢这本书。

我要感谢北卡罗莱纳大学为我的写作提供的时间和资源保证。感谢以下同仁：主任 Marcus Boggs、项目经理 Karol Jurado、助理编辑 Gina Kessler、编辑助理 Christina Ganim、技术项目经理 Lauren Keyes、市场经理 Karin Sandberg 以及市场助理 Natasha Coats。我还要感谢本版的审阅人提出的有用的建议，他们包括 Briar Cliff 大学的 Jennifer Bonds-Raacke 博士、Duke 大学的 Daniel Cerutti 博士、Milligan 学院的 Joy Drinnon 博士、Southern Indiana 大学的 Julie Evey 博士以及 Sierra 学院的 William Hardy 博士等。最后，我要感谢我课程中的学生，他们通过自己的成绩告诉我本书的成败，许多来自不同国家的学生在学术会议上见到我并告诉我，他们喜欢这本书。

David W. Martin
david_martin@ncsu.edu

目 录

1

如何开展有序的观察

伴随着疑问、想象或创造发明的直接的直觉观察只是一种有限的，甚至是误导的前科学方法。

C. F. MONTE(1975)

生命的本质在于，当你把它切开来仔细研究时，它就不是生命了。作为生命副产品的行为更是难以捉摸。

K. Z. LORENZ(1962)

在行为科学中，动物被试的反常表现必然会大幅度增加实验研究的复杂性。

S. N. ROCOE(1980)

本书的主要目的是教你如何做心理学研究。既然在许多大学的心理学专业中这些都是必须的技能,那么,为什么还要在这里学习这些内容呢？其中的一个原因是你想成为心理学家——一个研究人或动物行为的科学家。实验方法是收集数据、构建心理学知识体系的主要方法之一。本书也将简要介绍一些其他心理学研究手段,但重点还是放在如何做实验上。

即使你不想成为心理学家,学习一些心理学的实验方法也有助于使你成为一个受过良好教育的人,并为你提供一些可以用于其他职业领域的有用技能。例如,假设你进入了银行系统并凭借自己的努力成为了副总经理。显然,你学习的一些心理学课程对你有很多帮助,因为你了解一些人际关系的规律。实验研究方法同样对你有帮助。你的老板给你打电话说:"你知道,我们在银行中安装了语音自动提示系统,我们在这种新设备上花了很多钱,但不知为什么顾客并不喜欢使用它们。我想让你去弄清楚其中的原因,同时对它们作一些必要改进,以便让顾客能够喜欢使用它们。"

通过阅读这本书你会发现,要完成上述任务,需要许多做心理学实验的技能,虽然这并不是一个正式的心理学实验。首先,你必须作出假设:为什么客户不愿意使用这种自动语音提示系统？是他们觉得与机器对话使自己失去了个性？他们是被迫使用的吗？还是他们不了解如何使用这种机器？或者是他们觉得如果没有保安在场时在 ATM 机上存取钱不安全？其次,你要通过访谈或问卷的方法缩小假设的范围。然后,你可能想通过某种操作以改变客户的行为。如果知识普及有问题,那么,就提供培训;如果缺少使用动力,那么,就提供优惠的价格;如果安全有问题,则需要增加私密性。最后,你可能想测量客户的行为,以确定你的改进方法是否有效。尽管你的老板并没有要求你做一个心理学实验,但你确实在一步一步地按照心理学实验要求做心理学实验。许多工作都需要解决人的问题,本书中将为你提供这方面的技能,使你成为能够有效解决人的问题的人。

如果你的确想成为心理学家,那么,学习研究和实验方法的原因就显而易见。当然,如果你想成为一名实验心理学家,做心理学实验将是你主要的研究活动,你会反复使用本书中所教的方法。此外,即使你想成为一位临床医生或咨询师,至少你也应该了解如何做心理学实验,理想的情况是你能够亲自做这些实验。临床心理学家与其他从事治疗的治疗家(如社会工作者、心理咨询师等)的主要区别在于他们的工作与行为数据密切相关。在临床训练历史上(大约 50 年前),教育者们讨论决定临床心理学学生首先应该被训练成为科学家,然后,才是治疗家;如果没有科学上的训练,学生们将只能去猜测哪一种治疗方法起作用,哪一种不起作用。这就是为什么多数临床心理学家获得的是哲学博士(Ph. D, doctor of philosophy),这是一个研究性学位。当今,有四分之一的临床心理学家获得的是心理学博士(doctor of psychology)而不是哲学博士,但是,学位所要求的课程仍然要求学生在一定程度上精通研究方法。临床医生必须了解研究与实验过程,否则,他们就无法确定各种治疗的效果,也不能评估新治疗方案的合理性。

通过了解心理学实验的应用价值,我希望你愿意学习它,因为它非常有趣。我们对周围的世界非常好奇。我们想了解事物发生的规律。人类发明了科学,以便更

好地了解周围的世界①。科学是一种试图揭示这一过程的有序的方法。就我而言，在早期生活中，我发现，实验方法是科学研究的最重要工具，因为它能够揭示以前不为人所知的关系。此后，当我学习了心理科学后，我进一步发现，实验方法是一种强大的研究工具，它可以用来研究人们最感兴趣的课题——人类行为。

大多数人都对自己和他人的行为感兴趣。这就是为什么人们喜欢看肥皂剧、在背后议论别人、幻想以及在杂货店排队时阅读 *The National Enquirer*（美国新闻类杂志《国家询问者》——译者）等，这都是为了探究人的行为。心理学中的实验方法可以帮助检验我们的推测。在我的第一次实验心理学课程中看到以前从未见过的科学规律时，我是多么激动。即使在实验的若干年之后，在我看到一个新的实验结果时，仍然心跳加速。我的同事可能对我冲进他们的办公室向他们展示令人激动的实验结果已经习以为常。我希望你在做研究时也能有这种欣喜若狂的状态。尽管研究者从事心理科学研究的理由各不相同，但你也许一直对实验研究充满热情。

作为一门科学的心理学

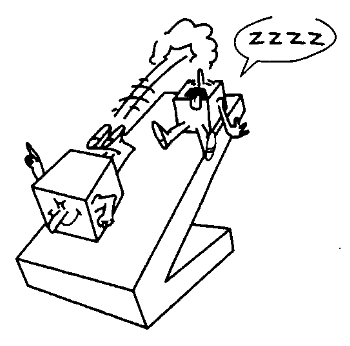
心理学家的工作与其他领域科学家的工作类似。在他们探索与理解人类行为的过程中，心理学家试图：①建立环境与行为之间的关系；②将这种关系纳入系统的知识体系。在本书中，我们将主要讨论第一个问题，第二个问题也将在第 3 章和第 13 章中有所涉及。

作为一名科学家，什么样的关系是可以接受的呢？当我们能够证明一个事件与另一个事件之间的关系在一定程度上是可以预测的，那么，我们就可以认为它可以被纳入科学知识体系。至少一个事件是可测量的行为，然而，行为的本质与其他科学有很大差异。就心理学家而言，我们主要关心的是人类行为（有时也包括动物行为）。这就是我们要讨论的第一个问题，即这个问题只有心理学家（而不是物理学家）要面对。人类和动物都是变量。我们人类不能很精确地重复一个反应，不论我们是否希望重复这种行为。在事物的可变性方面，物理学家的工作要比心理学家更容易。

物理学家可能采用测量一个木块滑过一个斜面所需要的时间来确定它的摩擦系数。尽管每次所测量到的时间可能都不相同，但这种变化相对较小。因此，虽然存在这种误差或只是测量一次，物理学家也不会出很大错误。然而，如果心理学家在测量人在灯亮时按键时间却不考虑个体差异，就会产生很大误差。物理学家的木块就不会因为分心、没有准备好或者眨眼、打瞌睡等这些人们经常有的行为而慢下来。

心理学家不仅要考虑不同试验之间的差异，同时还要考虑不同人之间的差异。物理学家可以制造出另一个大小、重量和表面完全相同的木块，以重复原来的试验。然而，心理学家却不能制造出一个人来。人的基因很少有完全相同的（同卵双生子除外），他们的生活环境也不尽相同。由于上述原因，一个人对灯光的最快反应速度可能比另一个人的最慢反应更慢。因此，作为心理学家，我们不得不面对同一个

① 天文学如此，其他世界也如此。



人的不同反应之间的差异，还要处理好人与人之间的差异①。

控制变异性的一种方法是使用统计技术。许多心理系学生通过学习课程体系中开设的统计课程学会这种技术。由于本书不是一本统计教材，因此，我不会花很多时间去讨论统计方法。只是会在第 12 章中提及这方面内容，而在附录 A 中将介绍一下简单的统计过程。控制变异性的另一个方法是尽可能在实验设计中对它们进行控制。本书的目的是帮助你做好实验研究，简单来说，就是："知道变异的来源，并且能够很好地解释它"。

与其他科学家一样，心理学家也使用一系列研究技术开展系统的心理学研究，从而解释这些变异。在本章中我将简要介绍这些技术。

本章和其他章节还将详细介绍实验方法，这是本书的重点内容。第 10 章将介绍一些非实验研究方法，如问卷法、单被试设计以及准实验设计等。

最为广泛使用的研究技术是定量化设计，在这种方法中，事件可以被量化或用数字表示。这些设计包括实验方法和相关性研究。为了使读者对研究技术有一个完整的了解，本章还将简单介绍质性（定性）设计，在这种方法中，事件很难被转化成数字。

量化设计

实验方法

作为科学家，我们就是要寻找事物之间的联系，但事物并不只是行为。事实上，我们在做一个实验或使用实验方法时，我们所研究的就是一系列不同条件与行为之间的关系。物理学家研究的是一个木块从一个角度、表面特性和温度都已知的斜坡上滑下所需的时间。而心理学家则不同，他们要研究的则是学生在班级中的行为。上述两个领域的科学家都试图探索一系列条件与行为之间的关系，这里的行为分别是自然物体和人的行为。这些关系是科学事实，它们是我们构建科学体系的材料。

然而，通过设计实验来揭示关系并不容易。在理想的情况下，我们能够穷尽所有的条件；然后，测量出各种条件下的所有行为。那么，我们就可以说，当某种情况出现时，一定会产生相应的行为。然而，如果我们能够列出所有情况，那么，我们就肯定可以找到一个用于描述这些条件的体系。仍然看上面的例子，如果我们想要研究学生在班级中的行为，那么，哪些外部条件是我们所感兴趣的呢？也许我们希望知道教师性别与着装、班级人数、教室中是否有计算机或者每天开班会的时间等因素的效应。正如你能够想到的那样，有许多因素都需要我们去探索。事实上，其中

---

① 你可能理解为什么心理学家用动物做实验。尽管他们能够在类似的环境中喂养遗传特性相似的动物，但如果在人身上这样做的话就会受到强烈的批评。你的朋友也许说："所有人都是动物。"或"所有女性都一样。"但是，千万不要相信他们。

的因素是无限的,它们构成了一个我们永远无法重复的环境结构。

与物理学家的工作类似,心理学家也试图揭示环境条件与行为之间的关系。当然,他们也会遇到同样问题。哪个行为是我们要去研究的呢?也许包括如何让学生注意力集中,或学生应该做多少笔记,或问学生多少问题,或学生的出勤率,或学生的脑电波,等等。同样,这里有无数行为都是我们可以选择的测量对象。

因此,科学家面临着无数的条件情景和无数行为,它们之间都有可能存在相互联系。选择行为要比选择情景条件容易得多。一旦选择了一个行为,就可以忽视其他行为。唯一的可能性就是只有在保持情景条件恒定的情况下,才能较为准确地确定情景条件与行为之间的关系。然而,如果我们这样做,那么,我们的结论也数不胜数。因为每一个情景条件都可以与无数的行为相对应。尽管条件与行为之间的关系非常精确,我们仍然无法预测在这种条件下将来会出现何种行为。因为在某一特定情景下不可能再一次出现某种特定行为。如何解决这一问题呢?

科学家不得不作了一个妥协。在他们选择了一个特定情景条件后,他们变化其他一些条件,使得这些条件至少在一定范围内发生变化。这意味着这个条件(或多个条件)被精确地确定下来,而其他条件构成了一个变化的而不是单一的集合。通过这种方法,所研究的情景条件与行为之间的关系通常可以推广到情景条件集合内的其他条件中。

科学家在使用实验方法时至少对一个情景条件进行操控,并且至少测量一种行为。例如,假设我们要研究词和图片哪一个更容易记忆。我们可能会制作一个词表,如汽车、树木、马和手等;然后,找出这些词的素描图。向人们呈现这些词或图,并比较他们分别要用多少次才能记住这些词和图片。我们将选择操控的条件分成两个水平,即词汇与图片;测量每种水平上记住这些材料所需要的次数。通过这种方法完成实验后,我们就可以得出一个清晰的结论,即采用呈现词和图片的方法对学习效果的影响。当然,我们不可能忽视所有其他条件。正如下一章要讨论的那样,我们必须认真考虑那些尚未操纵的条件因素。然而,如果实验过程正确,那么,我们就可以得到一个非常清晰的结论,即在操纵的情景条件中产生的行为是由操纵而引起的。实验方法在科学研究中得到了广泛使用,因为没有其他方法能够帮助我们作出如此可靠的因果推论。本章还会讨论其他方法,显然,所有其他方法都不是非常理想,因为它们都不能够非常确切地说“条件变化引起了行为改变”。

相关研究

在探索人类行为奥秘的过程中,实验方法未必处处行得通。在这种情况下,相关性研究通常更加有效。在相关性研究中,我们试图确定两个变量之间是否相互关联而不需要对任何一个变量进行操纵。例如,我们要研究儿童所居住区域的人口密度与该区域不良儿童比例之间的关系。我们的假设是,居住在高人口密度区域的人处于紧张状态,这种状态会使不良少年的比例增加。为了使得这个问题适合实验模式,我们必须将人口密度看成可操纵的条件,强迫街区交界处的婴儿家长居住到人口密度不同的社区。当儿童到 18 岁时,我们可能会去未成年人法庭计算每个儿童在这里出现的次数。很明显,不仅很少有家长愿意参与这样的实验,而且社会也不会赞同我们用这种方式开展科学研究。因此,为了能够继续研究这一重要问题,我

们可以考虑采用相关研究方法。

在开展这种研究中，我们可以从各种人口密度社区中随机选择一定数量的儿童。根据人口密度大小将它们用数字表示。然后，通过对法院中每个儿童的记录进行统计，以确定人口密度与儿童不良行为之间是否存在某种关系。

相关性研究的数据①可以用散点图描绘，图中每个变量分别在坐标上表示，每个点分别表示每一次测量值。例如，人口密度研究中的假设性数据如图 1-1 所示。图中的每个点表示人口密度分数和每个儿童在法庭中出现的次数。图右上部的点表示一个生活在高人口密度社区的儿童在法庭上出现了 4 次；而左下角的点则表示一个生活在低人口密度的儿童并没有在法庭中出现过。我们假设例子的散点图说明，人口密度与出庭次数之间存在中等强度关系。这些点似乎可以聚合成从左下至右上的一条直线。在这个例子中，生活在低人口密度社区中的儿童出庭次数更少。

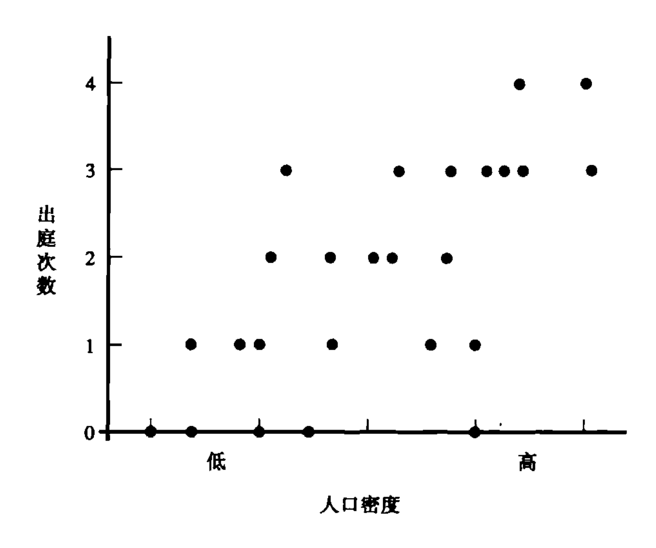


图 1-1 儿童的出庭次数和人口密度之间关系的模拟数据。

在本章开始时我们就提到，科学研究就是揭示事物之间的关系。为什么相关研究结果不如实验研究的结果那么完美呢？也许你还记得，在有关实验方法的讨论中我们曾经提到过，如果一个实验设计足够严密，那么，我们可以得出结论，即我们所操作的条件的变化引起了所测量的行为变化。而在相关研究中，我们能够得出的最好结论也只是两个变量是相关联的。

为什么我们不能说人口密度导致了青少年不良呢？这是因为我们没有对任何条件进行操作，而只是对两个变量进行了测量。只有在对人口密度这一变量进行操纵后才能作出因果推论。我们并没有强制不同家庭选择自己居住的社区。因此，选择何种人口密度社区居住是一种行为，而不是一个操作的条件，这并不符合用实验方法进行推论的条件。为什么呢？

就相关研究而言，一种行为可能引起了另一个行为。然而，即使是这样，我们也不能确定哪一个是因，哪一个是果。这个问题有时被称做“方向性问题”。例如，上

---

① 每个好的实验者都必须记住，数据(data)是一个复数；datum 是复数，data 也是。即使你每天早晨起床时对自己反复说三遍：“There data are(这些数据是)”，但你仍有可能忘记这一点。

述例子中的情况可能是，当儿童出现不良行为后，它可能会说服家长搬到人口密度高的社区居住。换句话说，青少年的不良行为引起了人口密度增加，尽管这种情况不可能出现。因此，在相关研究中，尽管变量之间存在因果关系，我们无法确定因果关系的方向。

另一种可能是，即使两个行为之间存在某种关系，但它们之间没有直接的因果关系。第三个变量可能引起了这两个变量的变化，这种情况被称做第三变量问题。在上述的例子中，贫困作为第三个变量可能对居住地选择和不良行为都产生了影响。

下面的例子可能说明相关研究中无法进行因果推论的原因。美国军方开展了一项摩托车事故方面的研究，试图揭示事故数量与社会经济水平、年龄等变量之间的关系。他们发现，最理想的预测变量是骑车人身上纹身的数量。显然，纹身引起摩托车事故或者事故引起纹身的推论都是错误的。第三个变量与其他两个变量都有关系，即冒险偏好。一个喜欢冒险的人喜欢在身上纹身，也可能在骑摩托车过程中更容易出现事故。



我相信，你一定知道烟草行业与政府在吸烟与健康之间关系的旷日持久的争论。几十年前，美国外科医生协会面临的两难困境就可以很好地描述用相关数据进行因果推论的困难。虽然人们已经发现吸烟数量与肺癌或其他健康问题之间存在正相关，但外科医生协会仍然不愿意作出吸烟导致肺癌的结论。也许其中存在着某种政治上的考虑。而这在科学研究中是合理的，因为也可能存在第三个变量同时影响着肺癌和吸烟。例如，易紧张的人体内产生的化学物质使他们的身体处于应激状态，应激状态下的细胞容易出现恶变。焦虑紧张的人也容易抽更多香烟。因此，紧张可能对两个变量都产生了影响。

因此，外科医生协会官方可能不得不展开实验，以得出吸烟导致肺癌的确切结论。这样的实验可能需要 1 000 人每天吸 40 支香烟，另外 1 000 人每天吸 30 支，等等。在这种设计中，实验者可以确定每组中的所有人在一生中患肺癌的可能性。如

果这个实验过程是合适的,那么,不同组中的个体患癌症的差异就可以归结为香烟。但是,我们的社会却坚持认为,个人的爱好应该得到尊重;这种实验存在着严重的伦理问题,它永远不会被允许。

那么,如何在烟盒上印上:“医师协会警告:吸烟会导致肺癌、心脏病、肺气肿,也可能导致难产等!”?①。在这种情况下,实验者确定其他许多变量也可能与健康问题和吸烟有关。随着一个个变量被排除,吸烟就可能是导致健康问题的原因。医师协会可能认为,实验者最终已将所有其他可能的变量都排除了。此外,结合动物实验,于是,就作出上述结论。

因此,有时我们也通过相关研究揭示重要的心理关系。然而,我们在作出结论时也必须非常谨慎,不要错误地将相关关系看成是因果关系。

获得相关数据的一个重要方法是调查法,它可以使用问卷或访谈的方法。由于学生在使用本书作为研究方法课本时,经常将问卷法作为课程实践内容。我在第10章中将具体讨论问卷法。这里只是简单介绍一下调查法。

调查法

调查法通常是去询问人们的行为和观点。你自己也可能参加过很多调查,有时甚至都没有意识到自己被调查。例如,在我们学院,毕业生会收到一封信,调查他们在大学的一些经历,包括教授授课的效果、医疗状况、是否容易获得职业咨询以及食物是否可口等;或者有人通过电话向你询问一些政治方面的问题,如对某一事件或候选人的看法等。在互联网上,为了获得某种服务,你会提交一些关于你是谁、你的喜好等问题的答案。这些都是调查。

调查法中的问卷通常是一种纸笔表格,它可以单个进行测试,也可以集体实施测试。问卷可以通过信件或互联网传递。调查法还包括访谈。访谈可以面对面进行,也可以通过电话进行。在第10章中我将详细讨论它们的优缺点。

调查法有不少优点。一个优点是,你可以直接询问被调查者的观点、态度、动机等,而不需要从他们的行为中进行推测。例如,我们可以用实验方法通过改变商店内货物摆放的方式来确定如何让顾客购买更多商品。然而,尽管我们发现顾客购买增加了,但我们仍然不知道其中的真正原因。也许是顾客对这个商店有好感,也许是某些商品更容易被看见。调查法可以告诉我们顾客购买数量增加的原因,至少可以了解他们自己觉得购买的原因。第二个优点是,它可以很容易获得大量数据。例如,我曾经看过总统的国情咨文的电视转播,就在转播后的几分钟之内,电视网络观众的调查结果就公布了。

调查法也存在一些缺点。尽管你认为被调查者作出的反应是真实的,但是,他们说的内容与实际情况之间仍然可能存在差异。例如,盖洛普公司对人们是否去教堂进行了60年的调查。他们发现,有40%的人每周都去一次教堂。这一数据比其他西方国家的数据高。很多教堂也发现,最近一些年来,教会人数在减少。到底哪个是真实情况呢? C. Kirk Hadaway 和 Penny Long Marler(1998)咨询了牧师并对教堂中的人进行逐个清点。他们发现,去教堂的人数比例在20%左右,而不是40%。

---

① 实际上,有很多关于吸烟损害健康的警告,但所有这些警告都暗示是吸烟导致了健康问题。

为什么这些去教堂的好人会说谎呢？原因可能是，有些人即使上一周没去教堂，但他们的确经常去，因此，就回答"是，每周都去！"。或者是，他们认为，好人应该去教堂，他们也想让自己被别人看成是好人，因此，回答说自己每周都去。无论什么原因，人们都是倾向于夸大自己投票或参加慈善事业的次数，而低估了自己使用药品或用办公室复印机为自己复印私人文件的次数。作为一个研究者，你必须记住，调查法的最大问题是调查结果只能说明，人们说他们如何做或如何想，而不是真实的如何做或如何想。

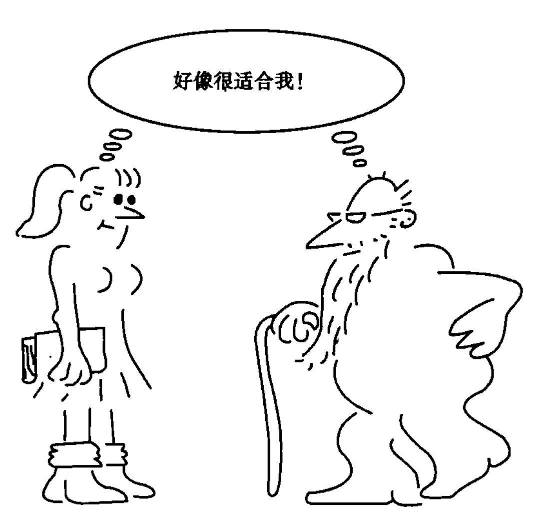
调查法的另一个不足之处与它的优点一样，即数据量太大。问题之一是采集这些数据需要大量被调查对象，而有些情况下被试是有限的。例如，在我们大学中，研究者以心理系学生为被试的研究必须等到其他研究者完成自己的研究一段时间之后才能开始，因为如果太密集的调查会使被试产生厌烦，从而影响研究结果的准确性。更为严重的缺点是研究者很难揭示如此大量的数据的内在规律。我曾经读过学生写的许多调查报告，在这些报告中他们列举出了调查结果，但更多的内容就不知道从何说起了。由于调查研究很少能够产生理论，它的结果无法支持或拒绝理论，而实验则可以做到这一点。另外，对调查数据的详细分析常常会用到一些复杂的统计技术，如因素分析，而新手并不熟悉这些技术。我将在第10章中更详细讨论调查法的优缺点，并详细介绍问卷调查过程。

档案研究法

相关研究的另一种方法是档案研究法。这种方法利用其他人的记录——即公共或他人的记录中可能包含了对你有用的信息。当你在研究中对这些记录进行组织分类并试图从中发现某种关系的时候，这就是在做档案研究。我把这种方法也归类为定量研究，这是因为大部分记录是可以被量化的。然而，如果这些记录中包括访谈、个案等时，这种研究就可以被看成是定性（质性）研究。心理学家所感兴趣的记录包括人口学数据、法庭记录、报纸、医院的文件、事故报告、犯罪报告、临床报告、政府文件、公务员的工资表、电话簿以及公司的销售记录等。

下面举一个档案法研究的例子。亚利桑那州立大学的 Doug Kenrick 和他的同事通过大量报纸上刊登的结婚情况的分析（Kenrick & Keefe, 1992），对人际吸引的进化理论进行了检验。进化理论认为，在远古时代，男人对女人的吸引力主要体现在男人是否具有为他们的孩子提供足够的生存资源的能力，而女人的吸引力主要表现在是否能够生更多的孩子。如果这个理论正确，那么，在现代的婚姻中，女性通常会嫁给年龄较大的男性，因为他们已经积累了较多的资源；而男性则会娶较为年轻的女性，因为她们的生育年限更长，可以生育更多孩子。为了检验这个假设，Kenrick 等对报纸上刊登的结婚者的年龄等进行了分析。正如进化理论所预测的那样，研究者发现，新郎年龄的确比新娘的年龄大。当然，这种年龄差异也可能存在其他原因，研究者也对此进行了详细讨论（Kenrick & Keefe, 1992）。因此，这项研究表明，即使是报纸这样公开的档案信息也可以作为产生有意义心理学研究的材料。

一个被最为广泛关注的例子是 Martin Daly & Margo Wilson（1988）的书中关于杀人行为根源的研究。研究者也采用进化理论解释这种行为。这个理论通常预测，如果一个人要去杀人，他的目标通常是那些生育成功可能性最小的个体；最不可能



的对象是他们的亲生子女，因为他们携有自己的基因，或者可以帮助他们现在或将来的子女成长的人，例如，信赖的配偶。因此，该理论认为，父母更容易杀害养子女而不是亲生子女；男性更可能因为配偶不忠而杀害她。研究者仔细分析了底特律市和加拿大的警察局中关于杀害案件的文件，他们发现，所有的预测都是正确的。养子女被杀害的可能性是亲生子女的40至100倍；性别嫉妒是男性杀害配偶的主要原因。

有些数据甚至可以支持与常识相反的预测结果。例如，对于一定年龄的成人子女而言，年龄较大的父母被子女杀害的可能性高于年龄较小父母被杀害的可能性。进化理论可以解释这种现象，即年龄较大的父母再生育小孩的可能性较小，而这些小孩可以延续他们的基因。但这一结果与许多心理学理论是矛盾的。Daly & Wilson 在研究中广泛使用档案研究法，他们将大部分资料转换成数字，并用统计方法进行分析。

档案研究法有许多优点。如果能够找到适当的资料，你就不必花时间和精力去收集数据了。另外，在有些情况下，现有资料所提供的数据也远比你自己获得的数据更加具有广泛性。最后，有些记录资料是在你自己的研究中无法获得的。心理学家不可能因为要作研究而鼓励人们相互残杀或相互随便结婚。当然，这种方法也存在一些缺点。与相关研究和自然观察类似，被试不是随机选择的，对所有变量的操作也都不是独立进行的。因此，获得的只是一种相关关系，而不是因果关系。此外，在多数情况下，那些收集记录的人并不是训练有素的科学家，因此，记录的可信度是未知的，甚至值得怀疑。有时，这些记录资料也难以被发现或得到；即使能够得到，也可能很难对它们进行系统化组织。最后，多数情况下，也不容易找到能够提供你所需信息的资料。

定性(质性)设计

心理学中的大多数研究者都使用定量化设计，如实验法或相关研究；这是因为在心理学历史上，心理学的自然科学特性满足了它作为硬科学(hard sciences)(如物理学、化学等)的条件。早期内省主义心理学家使用口语报告而不是可测量行为作为研究数据的短暂尝试被行为主义者击退，直到数十年后才重新出现。然而，近些年来，有些心理学家，尤其是在教育、临床以及社会等心理学领域也意识到这些死板规则的局限性。他们已经开始寻找那些可以将口语报告作为研究数据，同时又能够保证科学性的方法。他们已经从人类学、社会学中借用了一些方法，这些方法被称做定性(质性)研究。定性研究者使用描述性数据，例如，对人的描述(包括意见、态度)以及事件和环境的描述等。

民族志

假设一个文化人类学家长途跋涉到异国他乡开展研究,那么,这位人类学家将如何开展工作呢? 他或她对异国文化知之甚少,以至于无法设计实验去开展研究。即使编制出问卷或设计结构严密的访谈也非常困难,人类学家只有了解当地民众的基本情况后才能顺利开展研究工作。因此,开始的目标是与当地人交流,了解他们和他们的生活环境,以便熟悉他们的文化。有时定性研究中与此相反的方法就是民族志方法,即研究熟悉的文化,使它们变得陌生(Erickson,1973)。例如,我们要研究小学中一种教学方法的效果。我们去了所有的小学,对那里的情况非常熟悉。如果我们想从这些小学中获得新的信息,民族志学者的建议是,我们应该像外星人一样,第一次看到这些班级,把所有的事情都看成是陌生的。

我们应该从访谈学生和教师开始,用不带有任何偏见和假设的完全开放的心态去看待班级中发生的任何事情。我们不能随随便便地开展访谈。访谈中,不能只是依赖自己的记忆力,而是要将访谈内容录下来,并将它们及时转换成文本。我们可以通过笔记记录班级中个体的行为、发生的各种事件以及这些事件发生的环境和条件。由于民族志学者通常不会对他们的数据作出解释,因此,我们应该尽可能准确地描述学生和教师的言行以及班级环境。有时民族志学者也作为一个参与的观察者。例如,在班级研究的例子中,教师可能会作为一个参与观察者开展研究。参与观察者应当尽可能不要言语太多,以避免误导其他人的行为。例如,只在休息的时候才做笔记。

上述例子也可以被看成是自然观察,我们将在后面讨论这种方法。当然,并不是所有民族志或定性研究都在自然条件下进行。

我的一个同事对母女关系以及这种关系的演变过程开展研究。她采用的方法是在实验室访谈,既访谈母亲,也访谈女儿,并用录音机录下来。然后,再将内容转换成文本。她对数据进行解释,而不像民族志学者那样只是进行一般描述。尽管她对访谈进行了结构化设计,以便可以在不同的访谈中讨论同一个问题;但是,她并不要求访谈中的问题必须按统一顺序提问。她设计的访谈问题很灵活,不是一个简单的口头问卷。定性研究者认为,灵活性是这种方法的优势,即在与被访者的交流与沟通时必须让他们能够用自己的方式表达自己的经验、感觉以及态度。这些研究者认为,实验方法收集的数据高度结构化,只是检验了有限的假设,因此,它的人工痕迹和局限性太大,无法涵盖如此大量的数据。事实上,定性研究者一直持有一个基本信条,即定性研究具有很强的优越性,因为它具有人文取向。在定性研究中,参与者被当成人对待,他们的人性得到充分兼顾,而实验法在实验中则将参与者看成研究对象(被试)。

自然观察

如上所述,有些心理学家认为,在自然环境下研究人的行为是最好的研究,而用填问卷或设计实验可能会曲解被试的行为。例如,我们要研究酒精消费量与社会攻击性行为之间是否存在联系。我们可以设计一个实验来研究这个问题,实验中被试按照醉酒时喝的酒量进行分组。然后,让他们坐在一个房间中相互交流,同时观察

他们之间的攻击行为。在这种情景中饮酒者会产生怎样的攻击行为呢？他们可能会类似于在教堂中的聚会，而不是挤在酒吧里。

为了有效地回答这个问题，我们可能不得不去酒吧观察那里顾客的行为。这种心理学研究方法就是**自然观察**，因为这是在自然场所的观察①。当任何人工设置的实验情景都可能会干扰被试行为时，就需要采用自然观察法。例如，成人的出现会抑制儿童行为，尤其在陌生人出现时更是这样。我们假设，儿童在自己家里玩自己熟悉的玩具时的情况与他们在实验室中面对陌生玩具和陌生的心理学家时会有所不同。

很长时间以来，比较心理学家②和行为学家一直在试图揭示除了人类外，是否还有其他动物能够使用工具。自然观察可以提供这个问题的初步答案。在一开始，通过观察动物园中黑猩猩的结果表明，其他动物并不会使用工具。然而，研究者很快发现，动物园中的黑猩猩不会使用工具可能是因为动物园中没有可使用的工具。于是，研究者就给黑猩猩提供了一些工具，如老虎钳、螺丝刀等；他们发现黑猩猩仍然不会使用工具。最后，一个名叫 Jane Goodall 的研究者非常聪明，她带着黑猩猩进入了森林，并在那里与它们一起生活了许多年。有一天，她发现一只黑猩猩拿着一个树枝，用树叶将树皮磨去，并使它的长度合适；然后，用这个树枝伸进白蚁洞中将粘在棍子上的白蚁拿出来。虽然这个木棍不如人类的工具那么复杂，但研究者认为，它对于黑猩猩来说，仍然是一个合适的工具。最近的实验室研究也证实了黑猩猩能够使用工具。有些研究者也证实，许多其他动物（如鸟）也能够使用工具。如果没有自然观察，研究者仍然只是坐在动物园中观察，无论如何也发现不了动物能够使用工具。

除了心理学外，其他科学研究中也将自然观察作为主要研究方法，因为有些变量无法控制。例如，天文学家必须在非常自然的状态下观察宇宙。考古学家、古生物学家、民族志学者以及人类学家等也都是这样。自然观察法的局限性并不能阻止科学家揭示一些重要的科学规律，如进化。由于变量控制存在的问题，研究者常常使用自然观察法初步揭示心理学中的规律，并相信今后可以在实验室中进行更详细地研究。在这种意义上，自然观察是也一种有价值的方法。



自然观察的局限性也是显而易见的。由于研究者没有对正在观察的任何变量进行控制，某一个变量可能与被观察的变量一起发生着系统变化。在上述的酒吧观察的例子中，研究者可能会发现，喝酒越多，攻击性就越强。然而，研究可能没有注意到，随着晚礼服和饮酒的增加，酒吧里的捐款数量也随之增加。攻击行为可能与拥挤有关，或者酒吧男招待过于劳累，拿酒的次数减少。

① 自然观察法有时也被称为现场研究，因为研究者要深入现场收集信息。（如果我能够控制自己，少说话，那么，我也许能够对那些出色的农民产生偏见……）

② 比较心理学家不是在电视的商业广告中对 X 牌子和 Y 牌子产品进行比较。他们是对包括人类在内的不同种系动物行为的比较；他们主张我们的研究过于自我中心，人类只是动物世界中的一小部分。

攻击行为也可能与挫折相关。

因此，尽管自然观察在实际应用中有它的优点，它仍然由于缺乏控制而显得美中不足。与相关研究相似，研究者必须意识到其中可能存在的潜混淆变量，并且不要作因果推论。

个案研究

最后一个被心理学家使用的定性研究方法是**个案历史研究法**（case history）。个案历史研究法是对个案中具体细节进行分析的研究方法，个案通常是一个人的生活经历，同时也可以是一些突发事件，如核电站突然停止了运转等。许多临床心理学研究的数据都来自于个案，这可以追溯到弗洛伊德的临床个案报告。作为一种典型的定性研究方法，个案研究中的数据通常是口头的。假设你作为一个治疗者对一对有多重人格并存的双胞胎进行治疗，你也许对这对双胞胎产生双重人格的原因很感兴趣。你很快会发现，用实验来寻找其中的答案是徒劳的。即使你能够找到足够双胞胎作为被试，但从伦理上来讲，你不能让同时发作的双胞胎变成精神病患者，也不能让非同时发作的双胞胎精神上出问题。你也许会采用相关法进行研究。你可能会计算同时发作双胞胎的人格数目与他们儿童时期的紧张程度间的相关度。当然，你需要找到足够数量的具有双重人格的双胞胎作为研究对象。由于这个任务是无法完成的，因此，以每个数据点为基础的相关研究结果并没有什么意义①，你也不得不放弃这种方法。

剩下的唯一选择就是个案法，这种方法可以列举出双胞胎发展过程中的各种因素。首先，你花数小时去访谈这对双胞胎，去了解他们从出生到现在的生活历程。然后，访谈他们的亲戚、朋友，收集所有可以获得的他们在学习、医疗及心理等方面的各种记录。由于这些资料的量很大，因此，你应该根据需要选择有用的内容。个案研究也同样有其他研究方法所具有的各种风险，包括各种潜在的混淆变量以及无法进行因果推论。这种方法也有缺陷。一方面，研究者通常试图根据相关人员主观报告的事件对研究对象的过去进行重构。然而，已有研究表明，人们对过去事件的回忆往往是不准确的。有研究者发现，母亲在 6 个月至 1 年后，对怀孕和生产过程细节的记忆就不那么准确了。你可以想象，20 年后的记忆会如何。

个案法的第二个缺陷是研究者在选择所研究的事件时容易产生偏差。在我学习心理学课程时，老师要求我们从小说《罪与罚》里重要人物的生活事件中找出能够支持某一人格理论的证据。选择能够支持某一个人格理论的事件非常容易。然而，我发现其他同学从同一本书中找到了支持其他三个人格理论的生活事件。他们或者选择了不同的事件；或者选择的事件与我相同，却作出了不同的解释。即使在这本事件有限的书中，偏差在构建不同关系中起着非常重要的作用。难道研究者真的能够从一个人一生中无数的生活事件中找到支持自己所钟爱的理论吗？

到目前为止，已经出版了许多分析著名历史人物人格特征的书，如尼克松、肯尼迪、弗洛伊德等。虽然这些书引人入胜，但这种所谓的心理故事仍然避免不了个案

---

① 仅用单一数据点构建两个变量之间的关系非常困难。然而，用两个数据点就不那么困难了，因为你可以在他们之间画出两条直线。用两条线表示一种关系非常类似于“吹牛”，这看起来很容易做到，但没有人去关注它。

方法研究法所面临的危险。另外，作者用于支持自己理论的事件往往也是来源于公共媒体的第二手资料。因此，作者可能会更加远离客观事实。例如，一个学者认为尼克松是精神病患者，另一个学者则认为他是一个神经质的人。

在实验场景中也经常使用个案方法研究那些不经常发生的事件。例如，心理学家不可能用实验的方法研究空难原因。因此，研究者就会尽可能详细地重建事故发生前的各种事件。通过收集足够事故的关键事件，研究者期望能够验证他们关于事故原因的假设。然后，就可以在实验室验证事故原因。

个案法最广泛使用的领域之一是神经心理学。神经心理学家和神经科学家对大脑不同区域的功能非常感兴趣。揭示大脑不同区域功能的主要方法之一是将这部分脑组织捣毁，然后观察其行为的变化。对人类而言，损毁脑组织显然是一个伦理问题。因为大脑组织不能再生，所以，任何损毁都会造成永久性伤害。一种解决办法是寻找那些在事故或疾病中大脑不幸受伤的人作为研究对象。在心理学导论课程中，你可能还记得 Phineas Gage 案例，在一次矿山事故中一个金属棒穿过了他的大脑。这是研究者了解大脑工作过程的一个案例。目前，研究者收集了大量不同神经病变患者的行为。这些数据与其他研究结果（如动物研究）一起可以帮助我们理解大脑功能。但是，我们必须意识到，这些个案数据并不是来自实验室实验，因此，在进行因果推论时还必须非常小心。

个案研究的一个明显优点在于当只有一个或很少个案时，仍然可以使用它。它的另一个优点是它可以用于研究复杂的自然环境中的行为，而实验研究则只能在人工环境中研究人的简单行为。然而，如前所述，由于个案研究存在着一些缺陷，如从长时记忆中主观报告的内容很难是准确无误的；因此，我们应该审慎对待单个个案研究的结论。

定量与定性设计

不幸的是，习惯于使用定量研究或定性研究的研究者往往会认为其他研究方法会误导研究。定量研究者认为，只有将资料转换成数字，它们才能够用于构建科学知识体系；如果我们不能够建立有助于理解人类行为的理论，科学就不会发展。这些理论建立在对行为原因了解的基础上；如果没有实验研究，仅靠一些相关研究是无法得出因果结论的。这些问题的根本在于数据的信度。没有重复的研究，我们就无法判断数据的可靠性。有些实验者认为，定性研究者，如民族志学者，只是在描述人的行为，他们所做的只是历史学家和作家的工作，不是一个科学家的工作。

而定性研究者则持不同观点。他们认为，实验研究只是在人工环境下探讨极小部分缺少人类属性的行为，定量研究者无法从整体上了解人的真实行为。另外，只有定性研究才能更接近揭示个体的潜质，通过研究者的努力和创造力构建我们的科学体系。甚至有定性研究者认为，实验研究将人当成被研究的物，而不是人。有些更加极端的定性研究者完全否定传统的科学研究，并坚持认为，定量研究者不愿意接受定性研究设计是为了保持某种政治势力和传统势力的压制。

一种合理和中庸的观点应该是只要能够解决我们的问题，任何一种方法都可以使用。定性研究至少可以为我们提出假设提供手段，以便用定量方法检验这些假

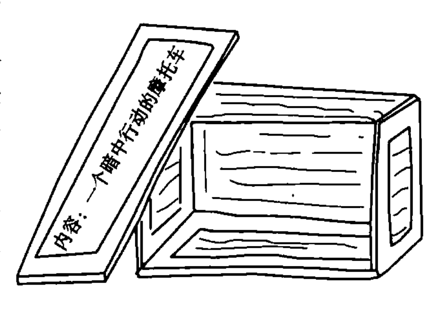
设。在有些情况下,我们也没有理由否定两种方法的结合使用。例如,许多调查即包含了定量内容,如使用里克特量表收集数量化信息,也包含了如开放问题这样的定性内容。在这种情况下,我们可以用定性数据解释定量化结果。下面的例子说明了两种方法相结合的优势。

不同方法相结合

为了说明如何使用不同方法去验证研究假设,请看下面例子。当你开车准备进入高速公路时,你会快速地左右观看,并开始加速;但突然停下,并对自己说,"噢,前面有一辆摩托车开过来,我差点没看到。"或者你是摩托车驾驶员,一辆汽车突然经过你身边,似乎没有看到你。为什么会出现这种情况呢?正如我们将在第3章中讨论的那样,我们在日常生活中遇到的这些问题可以帮助我们提出心理学假设。形成假设的第一步是检验不同情景相互联系。摩托车与其他机动车(如小汽车或卡车)之间存在怎样差别呢?很明显,摩托车比较小,与较大的机动车相比,更不容易被看到。但我们并不是第一个发现这种差异的人。当我们阅读了第6章后,你就会知道如何了解其他人是否已经对这个问题做了研究。密西根大学交通研究中心的Paul Olson曾经对这一研究进行了综述,并把我们的假设称做摩托车易见性假设(Olson, 1989)。我将以他引用的一些研究为例子,说明检验假设的研究方法。

首先,虽然我们已经知道个案研究存在许多缺陷,但它仍然可以帮助我们形成假设。为了验证摩托车易见性假设,我们使用的个案研究与传统的单个案研究并不相同;我们可能会发现那些闯入摩托车道的人,并询问他们为什么会这样。是否有一种方法能够更加系统地收集这些资料呢?幸运的是,警察已经为我们做了一些工作。在一定程度上,事故报告就是一个简短的个案研究。如果你收集摩托车事故的报告并仔细阅读,那么,你会发现,那些侵占了摩托车路权的驾驶员都声称自己根本没有看见摩托车或者没有及时看见摩托车造成了碰撞。如果易见性假设正确,那么,你就可以听到事故司机的上述描述。在解释这些结果时,我们必须注意到个案研究的一些局限性。

在上述例子中,尽管增加个案的数量可以增强我们的信心,但我们必须记住,数据依赖于人的记忆,而收集这些信息的并非专业人员,而是那些事故中的驾驶员,他们的反应会受到法律的影响。



我们是否可以用自然观察的方法验证我们的假设呢?如果我们有足够时间,那么,我们可以坐在街头,等待摩托车事故的出现并观察它的过程。当然,事故报告中有目击者或警察对事故的描述,这些信息可以帮助我们还原事故。我们也可以开展档案研究,查看不同类型事故的统计结果,并将它们与汽车相撞事故相比较,以确定两者之间的差异。如果我们这么做就会发现,一般情况下,小汽车和摩托车产生同类事故的频次相同,而在摩托车直行或小汽车在摩托车前左转的情形则是例外。我们也必须注意到,这种自然观察充其量也只是相关性研

究。实验者没有做任何操纵，只是测量行为的变化。我们也许可以将这些统计结果作为支持易见性假设的证据，但它的强度还不够。为什么在有些情况下摩托车不容易被看见？也许不论摩托车是否容易被看见，其他机动车驾驶员都看不见它们。因为他们在左转时只是向左看，而没有看前方驶来的摩托车。

我们能否用实验方法检验上述假设呢？在第10章中我们将讨论当无法用实验方法进行研究时，可以采用准实验设计方法开展研究。例如，在上述例子中，我们在验证摩托车易见性假设之前先分析一下每年的事故情况。幸运地是，自1967年起，大部分州开始要求摩托车在白天也要打开前灯。如果你测量某一特定行为(如摩托车事故)在事件发生前后的次数，这样就可以形成了一个准实验设计的间断时间序列。这种设计不是严格的实验设计，却比相关研究更加严密。有些研究者使用这种方法预测在白天摩托车事故率将降低4%到20%。然而，最近的研究表明，摩托车白天开灯措施对撞车事故的改善非常小，甚至没有任何改善。如果这种效应存在，那么，白天开灯措施也同样可以减少小轿车的撞车事故，尽管小轿车没有易见性问题。因此，车灯的研究结果是不确定的。

我们也可以设计一个正式的实验以揭示小汽车驾驶员对摩托车的行为表现。在一项研究中，研究者按照“2秒规则”估计与100米外迎面开来的汽车相会的时间。在这方面，摩托车、小汽车和大卡车之间不存在差异。而差异主要表现在机动车与迎面车辆安全避让的最后时刻，这个时间间隔对摩托车而言更短。然而，到目前为止，还没有实验能够验证易见性假设。

我分析了Olson(1989)的研究，其中包含明确地描述了各种可以检验单个假设的研究；同时，也描述了各种方法的优缺点。个案法和相关研究法相对比较真实，但不够严密和精确。相比之下，正式的实验研究更加严格，但不够真实。表1-1中描述了不同方法的优缺点。

表1-1 不同研究方法的优缺点

设计方法	优　点	缺　点
实验方法	控制精确、可以因果推论、测量结果精确、能够检验理论	人工环境、易受干扰、复杂行为难以测量、难以解释非结构研究
相关研究	可以发现变量之间的关系、通常可以进行精确测量、受到的干扰较少	无法得到因果结论、变量难以控制、需要很多被试
档案法	不需要更多的数据、可以研究那些罕见的行为、可以研究不可控的事件	难以得到适当的资料、数据不是由科学家收集的、数据只能进行相关分析
民族志法	可以描述不熟悉的情境、可以描述复杂行为、受干扰比较少、被试比较自然	对变量没有控制、难以精确测量、研究者容易产生偏向、不可能进行因果推论
自然观察法	真实环境有助于结果的推广、很少受到干扰	没有对变量进行控制、数据收集比较困难、研究者的偏向、无法进行因果推论
个案研究	可以研究稀少案例、可以对复杂行为进行深入研究	无法控制变量、数据依赖于对以前事件的记忆、研究者偏向、无法进行因果推论

我并不想让你无所适从。Olson 是怎样分析的呢？他认为，易见性假设还缺乏证据。最有可能的原因是由于摩托车比较小，很容易被其他物体挡住，如其他汽车、挡风玻璃上的支架或树木等。因此，驾驶员看不见摩托车不是因为摩托车不容易被看见，而是它被挡住了。

小　结

作为一个研究人类行为的科学家，心理学家可以采用很多研究设计，所有这些方法都是为了揭示不同事件之间的关系，并将它们纳入系统的知识体系中。实验法是一种定量设计，它是本书要重点介绍的内容。这种方法需要对环境进行控制，同时测量不同的行为。从实验研究中我们可以确定操纵的变量是否是某一行为的结果。

有时，当实验无法实施时，就可以用相关研究法。这种方法是对观察的变量之间关系进行分析，因为研究者没有对任何变量进行控制。相关研究常常采用量表或访谈方式开展研究；也可以通过档案研究中的数据进行相关分析，包括公开或隐私的文件记录，如人口普查或法庭记录等。

也有一些研究者采用定性法开展研究。定性研究者使用的是描述性数据，包括人们的书面记录材料（意见、态度），或者对事件或环境的记录等。在民族志方法中，研究者通过访谈或观察收集资料；或采用自然观察法从真实场景中收集资料。此外，当上述方法遇到困难时，个案研究也是有效方法，它可以用来分析个人的生活事件或历史事件。

2

怎样做实验

在直到19世纪的漫长历史中，许多有才能的思想家才开创了心理学，他们并未认识到通过仔细观察所获得的事实的重要性……最终心理学家决定，他们必须以物理学、化学和生理学为引领，将心理学变为一门实验科学。

R. S. WOODWORTH

我们必须防止……从实验和观察中预先作出结论。事实上，因为实验和观察的条件模糊，它可能会有多种不同推论。

K. DUNLAP

在第1章中我们简要讨论了实验法。你可能还记得，这种类型的研究最主要的好处是它允许你作出因果结论，即一种条件导致行为的某种变化。因为这种结论很精确，支持结论所需的规则都相当严格。大部分规则包含能够解释所有变化的条件。

例如，假设我们对人们根据特定强度的灯光作出按压按钮反应所需时间这一问题感兴趣。我们可以选择并操纵一个条件——灯光的强度，同时测量一个行为——按压按钮需要的时间。这两个变量都有正式的名称。

变 量

自变量

自变量是实验中最感兴趣的条件情景，如上例中的光强度。记住这个名称最好的方法是联想到这个变量不依赖于被试行为。

作为实验者，我们操纵这个变量，即选择两个或更多水平呈现给被试，同时被试不能改变我们已经选择的这些水平。例如，如果自变量是灯光强度，那么，我们可以选择一个高强度灯光和一个低强度灯光作为两个水平，同时观察两种条件下被试的行为。我们至少需要两个水平，否则，就不能做实验。我们可以选择更多水平或设置一个以上的自变量。在后面的章节中，还将讨论设计这些更为复杂的实验方法。

因变量

一旦选定了自变量，我们就会测量被试对这些变量的反应。那些用来测量被试行为的变量就是因变量，因为它依赖于被试做了什么①。例如，在反应时实验(reaction-time experiment)中，我们的目的是发现光强度和反应时间是否存在一定的关系。因此，因变量是从出现灯光到被试按键的时间。有时候对这种关系预先做一些陈述是有用的；这种陈述叫做假设。在这个例子中，我们可能假设：灯光越强，被试反应越快。实验的结果将会决定假设是否能够得到支持并成为科学知识体系的一部分。

我在这本书的许多不同地方都谈到了假设。在下一章中，我们将会讨论怎样从理论中推导出假设以及如果理论正确则假设必定是对的。在第12章中我们会谈到零假设的概念。众所周知，零假设只是一个统计上的陈述，它说的是从总体上来讲自变量对因变量无影响。但是，如果你真的认为你的实验里没有任何效应，你很有可能不会进行这项实验。事实上，你通常会预测自变量水平的变化将导致因变量的变化。这种预测才是你真正的假设。因此，实验者经常超出这种简单预测一些无方向性变化的假设，他们会进行一些方向性的假设，即当操作自变量时因变量变化的方向。

在一些实例中，假设甚至不是基于理论，特别是当你仅仅想知道当操作自变量

---

① 虽然“依赖”一词的确指的是行为是可能依赖于自变量的不同水平，但我认为用这种方式比较容易记住这一术语。

时行为会随之发生怎样变化。在本例中，假设仅仅是对问题的回答。例如，拥挤如何影响攻击行为？做一个猜想或思考较长的时间能导致多重选择测验中较好的成绩吗？那些在竞选海报上微笑的政客比那些没有微笑的政客更有可能或更没有可能赢得选举吗？对像这样一类问题的假设回答可以纳入知识的科学体系。

控制变量

至此，我们已经选择一个条件作为操作变量，即自变量。但在某些方面，我们需要解释实验中的其他条件。一种可能性是去控制其他的条件，因此，称它们为控制变量。我们可以通过保证它们处于一个不发生变化的单一水平。例如，在反应时实验中，我们可能需要恒定的室内灯光条件、右利手被试、常温等。理想的情况是，在实验中，除自变量以外的所有条件始终能保持在一个平稳水平上。通过这种控制，我们就可以知道因变量的任何改变肯定是由自变量变化引起的。

控制的概念对实验法至关重要，它使得实验法不同于上一章中介绍的其他研究方法。在实验中，许多变量将被设置为控制变量。实验者应当确信实验中控制变量完全得到了控制。这就是为什么心理学家们要花相当大的代价来建立一个声音、灯光和温度被控制的特别环境，用特别的设备来保证刺激特征的一致性，保证反应是被真实记录的。

但是，尽管你的实验中许多变量是控制变量，你还是应该意识到并非所有变量都会被设置成控制变量，尤其在心理学领域中。首先，实验者不能控制所有变量。这不仅是因为控制许多遗传和环境条件是不可能的；而且对于人类被试而言，也无法做到使合作态度、注意状态、新陈代谢速度以及其他许多情境因素保持恒定。

其次，我们实际上不想控制实验中的所有变量，否则，我们所创造环境就会过于独特。如果我们能控制所有变量并操纵自变量，那么，实验所建立起来的关系将仅存于一种特殊条件中，即所有的变量都精确设置在控制的那个水平。另一方面，我们也不能将结果推广到任何其他情境中。作为一个经验法则，实验控制得越严格，结果的适用性程度越低。

例如，如果美国空军的指挥官问你："我知道你进行了一项反应时实验。请告诉我应该把战斗机中的火警灯光强度调到多少时，我的飞行员们就能在半秒钟内作出反应？"已经进行了一个很好控制实验的你回答到："先生，如果你保证那个飞行员是 19 岁的大一学生，智商为 115，他坐在一个大小为 10 英尺 × 15 英尺的空调房间内，没有任何干扰声音，也不需要做其他事情；另外，还要你总是在灯光出现前 1 秒给一个提示信号，这样我可能能够给你一个答案。"你可以想象这位指挥官的反应。这个故事的寓意就是：如果你想要推广你的实验结果，那么，就不要控制所有的变量。

实验结果的推广也称为外部效度，即一个因果关系推广到不同的人、环境和时间的程度。Cook 和 Campbell（1979）曾定义出多种类型效度。他们通过这种方法认为，效度是实验中得出的因果结论是合理。我将会在本书的其他地方介绍有关专门用语。如果你想要将结论推广到不论年龄和智力水平的所有人群（包括我们上例中的空军飞行员），那么，当样本有限时，如只有大一学生参与实验，它的外部效度就会受到威胁。又或者你试图将一个高度控制的实验结果推广到真实吵闹的、热

且拥挤的环境和疲惫、动机缺乏又训练有素的工人群体中，外部效度也会受到威胁。总的来说，实验中控制得越多，即你选择控制的变量也越多，外部效度就更有可能受到威胁。

随机化变量

如前所述，我们不想控制所有条件，那么，我们还能够对实验中的变量做些什么呢？一种可能是让它们变化。在哪种情况下我们可以允许这些条件变化并仍保证它们不会使实验产生偏差？一种选择是允许一些条件随机变化。这些变量被称为随机变量。

随机或随机化被广泛应用在科学的不同领域中。它是从总体中随机选择一些个体，从而组成具有代表性的样本。在这个意义上，总体是可获得的；同时通过一些随机过程可以使总体中任何一个项目被选中的可能性相同。随机选择可以保证外部效度，即保证从总体中随机抽取的样本能够推广到整个总体。因此，如果你想要将实验结果推广到所有美国人，理想方法是运用一种平等的选择手段，即将这个国家所有人的名字放到一个巨大的帽子里，然后，选择一定数量的名字构成样本。这样你就可以说你是随机选择被试，并宣布你的实验结果有很好的外部效度。

但是，在“随机变量”中的随机一词通常指自变量水平的随机安排。实验室的许多条件与被试个体差异有关。很显然，如果同一批被试接受自变量的不同水平，那么，我们不必担心个体差异。但是，如果自变量的每一个水平上的被试都不同，那么，我们必须确保分配到每一水平上的被试特征不会造成结论的偏差。例如，假设你想要考察电视暴力对儿童攻击行为的影响效应。在你从一个较大群体中随机选择 200 个 6 岁儿童作为样本后，你可能要随机安排他们到自变量的 2 个水平中，即观看暴力电视节目和观看非暴力电视节目。你可能会通过抛硬币的方法确定每一个孩子分到哪一组；如果硬币是头像则将孩子安排在第一组，反之则安排在第二组。第一组儿童是否就读于充斥暴力的学校或吃糖太多，抑或他们的家庭具有暴力倾向，而第二组儿童却很少这样呢？当然，如果选择是随机进行的，那么，从统计学上来讲这种大样本是不可能产生偏差的。

假设你让儿童在家看暴力或非暴力电视节目。第一组大部分儿童在家拥有大屏幕影院系统似的电视，而第二组大部分儿童在家看的是简易的电视，可能会出现这种情况吗？再强调一次，这种现象是有可能的，但可能性不大，随机化使得这种可能性几乎为零。

随机设计和随机选择没有什么特殊技巧。你可以使用任何允许每一个体以等同概率进入样本的设备。如在上例中，如果你试图构成两组，你可以通过抛硬币来实现①。如果有六个小组，你可以掷骰子。如果有 33 个小组，你可以运用 33 张相同大小的纸片。大部分数学手册和许多统计课本都附有随机数字表。它是通过一种等概率处理过程，从 10 000 张纸片中抽取的随机数。在本书后面也附有随机数表，见附录 C。通过使用随机数表中任一行或几行，你可以给每个项目指定一个数字，

① 事实上，大部分硬币都会轻微偏向头像那面。但是，除非你的实验中的试验次数超过 10 000 次，否则，不用担心它。

并在这个数字出现时选择该项目；而忽略那些不在你的表上的数字。如果你精通电脑，你可以利用它产生随机数字或事件①。

如果你已经选择一种条件作为随机变量，那么，你必须确保它真的是以随机的方式变化的，因为并非所有看起来随机的事件都是随机产生的。例如，如果你试图通过自己指定方法将事件随机安排到不同实验条件中，那么，你并未做到真正随机化！人类创造随机事件的能力很差。如果你主观认为，被试整天或整个学期参加实验是随机顺序，那么，你就错了！那些上午与下午招募来的志愿者、上半学期或下半学期招募来的志愿者具有不同特征。新手实验者经常在随机化上犯错误。你难道没有那样做吗？

在实验中大部分随机变量可能与被试有关，并且可以通过随机安排被试实现随机化。但是，其他与被试无关的条件有时也可以被当成随机变量。假设在我们的电视暴力实验中，儿童观看电视的房间要么上午可用，要么下午可用。如果你认为上午或下午看电视导致儿童攻击行为差异的原因与攻击数量无关，那么，你可能试图随机地将暴力电视和非暴力电视组安排到上午和下午观看。你肯定不希望一个组只在上午观看，而另一组在下午观看。

还可能存在一些影响儿童攻击性的其他条件，尽管它们并没有得到真正随机分配，但你仍然可能把它们看成是随机变量。例如，雨雪天气、某一天新闻中暴力内容的数量等。因此，如果实验包括多个阶段，并且假定这些情境在自变量各个水平上随机分布以及它们不会对结论产生系统误差，那么，实验很可能是错误的。

如前所述，随机选择的主要优点在于结果的普遍性。每当你选择将一个情境作为控制变量时，你只能将结果推广到该变量的这一个水平。但是，如果在总体中某一条件存在多个水平，并且随机选择样本，那么，你可以将结果推广到整个总体。随机安排的主要作用是消除结果偏差。因此，随机化是一个强有力的实验工具。

限制随机

在有些情况下，你可能不想将某一条件变成随机或控制变量。事实上，随机化和控制确定了一个连续体相对应的两端。这两端之间是各种程度不同的随机化。在这种情况下，控制事件安排的一部分并对其他事件进行随机化。例如，在反应时实验中，练习可能是一个比较重要的变量。

如果我们在实验一开始总是呈现低强度刺激，紧接着全部是高强度刺激的试验，那么，别人可能会认为实验出现了偏差。事实上，对不同强度灯光反应的任何差异都可归因于练习时间的长短。为了避免这一问题，我们可以控制练习变量并仅仅给每个被试一次试验；或者随机安排低强度刺激和高强度刺激的试验，利用抛硬币来决定 12 个实验的呈现，出现头像时呈现高强度灯光，反之呈现低强度灯光。但是，这一选择可能不是最有说服力的，因为它可能导致高低强度刺激的不适当呈现。（例如，抛硬币使得仅有 3 个高强度试验和 9 个低强度试验。）为了避免这种可能性，我们会试图使高强度和低强度刺激的试验次数相等。

---

① 电脑在生成随机事件时不是完美的，但是这种方法比扔硬币要好。就实验中的事件分派而言，你所用的方法不会造成差异。

因此，作为一种解决方法，我们在试验顺序和次数的安排上（每种类型的试验拥有相同的数量）形成一种限制，并在这种限制内进行随机化分配。我们可能在6张纸片上写上“高”字，也同样将“低”写在另外6张纸片上；然后，从一个帽子中抽取纸片，以决定呈现顺序。这个程序可能满足在限制范围内不同试验在各条件下的顺序随机化，以满足两种强度平等出现的需要。

当然，也可能还有其他限制。我们可能要避免某一特定强度的试验出现在序列的开始部分的可能性。我们可以在区组内随机化，此时，区组就是限制。我们可以用这种方法选择3个区组，每个区组有4个试验，以确保在每一个区组内随机选择了2个高强度试验和2个低强度试验。为了描述这一过程，我们可能会说，将实验条件随机分配到含有4个试验的3个区组中，所采用的限制就是在每个区组中代表每一种强度试验次数相同。

只要你能够将这些限制具体化，你就可以合理地运用这些限制。但是，具体化的限制越多，选择过程的随机化程度就越低，结果推广的范围也就越小。

混淆变量

如果我们要设计一个完美的实验，能够选择操纵的自变量和测量的因变量，并将其他变量作为控制变量、随机变量或限制内随机变量，那么，我们就不必担心接下来要讨论的变量。但是，不是每个实验都是完美设计的。在许多真实环境中，设计一个完美的实验是不可能的。在这种情况下我们需要知道什么时候混淆变量构成一种威胁。任何随着实验者操纵自变量而发生系统变化的条件都是混淆变量。

例如，我们在反应时实验中运用了3个不同的灯光强度：前面20次试验为低强度灯光，然后是20次中等强度灯光，最后20次采用高强度灯光。如果我们说：“灯光越强，人们反应越快。”其他人可能说：“不是，人们在练习后反应更快。”实际上，上述两种说法可能都正确，或者其中一个是错误的！问题是我们无意混淆了实验，因为存在一个随自变量发生系统变化的变量。

一个实验者能记录最复杂的测量，做最好的统计检验，用海明威式的风格写出结论。然而，一个混淆变量能使所有努力变得毫无价值。可口可乐和百事可乐的长期争斗说明这种变量能导致一些混淆（“可口百事争夺战”，1976）。百事可乐公司与可口可乐公司在一次饮品测试中互不相让。在这项测试中，被试分别品尝可口可乐（杯子上标有字母Q）和百事可乐（杯子上标有字母M）。结果表明，一半以上的被试报告他们喜欢百事可乐。而可口可乐公司则通过他们自己的喜好测试加以反驳，即上述测试不是关于可乐的爱好，而是对字母的偏好。他们认为，选择杯子M的人多于选择杯子Q的人，因此，不是因为他们喜欢杯子M中的可乐，而是因为与字母Q相比，他们更喜欢字母M。这一假设被进一步实验证实，因为当两个杯子都装有可口可乐时，大部分被试仍然说更喜欢杯子M中的可乐。

在这个例子中，字母显然是一个混淆变量。因为在原先的测试中，他们随可乐类型发生系统性变化，实验者不能区分出品尝者的喜好是源自于可乐还是字母。

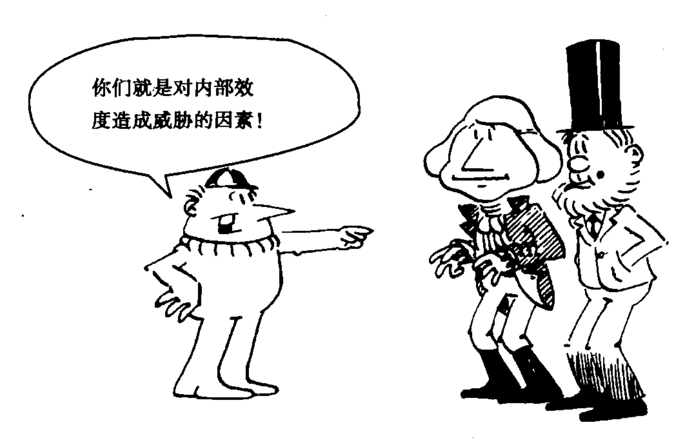
在本章开始曾提到，Cook 和 Campbell（1979）将效度区分为许多不同类型。另外一种效度就是内部效度，它指的是因变量的变化是否源于对自变量的操作，或是否有其他变量导致这种变化。如果自变量没有造成因变量变化，那么，混淆变量可

能导致这种变化。因此,如果我们想要在实验中避免出现混淆变量,我们需要了解那些可能对内部效度产生威胁的因素。实验者识别并避免(如果可能的话)那些在实验中威胁内部效度的混淆变量是非常重要的。

内部效度的影响因素

历 史

在实验室实验中,通常可以在较短时间内收集到自变量所有水平的数据。因此,因变量的任何变化都不可能被归因为历史因素,即那些发生在自变量不同水平测试之间的事件。



例如,假设你想要了解在大班授课的心理学导论课程上,使用计算机自动生成视觉投影幻灯是否比传统的手绘投影胶片更能提高学生的成绩;此外,某教授每年只开设一次这门课程。为了开展这项研究,你打算请这位教授今年用计算机自动生成视觉投影辅助教学,然后,把这个班的成绩与以前班级的成绩进行比较。如果今年的整体分数更高,那么,成绩的提高可以归因于计算机辅助的显示的作用。但是,一些历史事件也可能会导致这些变化。例如,学校可能已经加强了管理标准,因此,改变了班级中学生的学业质量;或者工程学院决定让所有高年级学生都修这门课,这同样改变了班级组成;或者由于人们对该学科的兴趣增强,就像个人电脑出现改变了人们对计算机科学课程的兴趣也发生了变化一样;或者在局部范围内,某些人获得了一份去年试题,班里也有一些学生获得这些试题。为了确定两个班级分数间的差异是否能够归因于使用了计算机辅助教学,你必须排除上述那些历史事件以及其他可能影响结论内部效度的因素。

成 熟

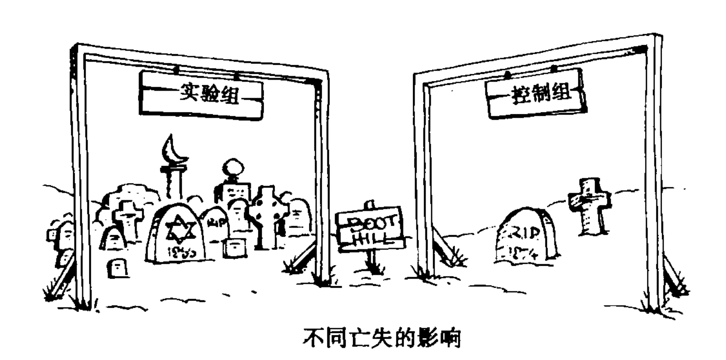
成熟是因被试成长或经验对内部效度产生影响的因素。与成年人相比,成熟显然对年轻人来说更是一种威胁。例如,评估学前教育方案的影响。但是,即使是成年人,在长期实验中或者被试在经历着快速变化时,成熟仍然是一个问题,如一个雇员首次开始负责管理工作。

选 择

当实验者不能随机安排被试,特别是实际操作中,当被试自己决定参加某些条件组时,选择可能会成为威胁。如果在上述例子中,选择作为比较的班级是秋季和春季班级,那么,选择就是一个问题。秋季心理导论课程班级包含许多大一学生,他们中的大部分是心理学专业的学生;而春季班有多数是学工程学的学生,他们在大四时才修这门课。这两个班级除了使用计算机辅助教学效果的不同外,还可能存在其他差异吗?实验者应该很清楚这一点,即大学生被试在事件上存在学期不同时段之间潜在差异①。总的来说,学期前半部分的志愿者参与实验动机更强,对实验充满渴望,他们很有可能都是较为优秀的学生,至少他们在计划时间方面的能力更强。影响最大的选择因素是那些直接与自变量相关的因素。例如,假设你想要评价一个新的工作培训方案。你让工人自愿参与这一新方案,然后在完成方案后比较那些志愿参加方案的工人与不愿意参加新方案的工人的表现,你是否认为自己选择的这两组被试可能存在差异呢?采用一种新疗法患者的康复率与拒绝该疗法患者的康复率差异如何?

亡失率

亡失率②及被试退出实验可能同样是影响内部效度的一个威胁因素。值得庆幸的是,在大多数的实验中,这些流失的被试仅仅是结束了他们在实验中的生命,而不是一般意义上的生命。总亡失率不是真正的问题;差别亡失率才是关键。它是指当很多或不同类型的被试从实验中的各个自变量水平组中退出。例如,假设一个公司决定尝试一个新的训练项目,从而为那些新提拔的中层管理者提供应对社会压力的方式。



不同亡失的影响

公司随机选取一半新经理,并让他们参与每天 1 小时与员工对峙模拟情景。其他经理则没有接受这种训练。训练后的 5 年间计算两组经理在对抗压力过程中的抱怨数量。公司发现,压力预防组的抱怨更少,因此,总结报告认为,方案是成功的。

① 实验中的志愿者类似于军队志愿者。虽然有些积极填答实验报名表的学生即使不是在课程要求的情况下也会自愿参加实验,但是大部分志愿者做这些都是出于一些其他需要,如写论文。

② 亡失率是 Cook 和 Campbell (1979) 用的术语。一些实验者也称它为流失率。

果真如此吗？人们必须问的一个问题是：训练期间，每组有多少经理退出①？相反，训练可能使经理更敏感，从而对与压力有关的健康问题意识更强。另一种有可能是，不仅会有更多经理退出压力组，而且这些经理可能对压力最为敏感。训练组的成功可能与预防程序几乎没有关系，但它可能完全归因于这样一个事实，即亡失率改变了小组的特性。

测　验

测验对行为的改变可以不受任何其他操作影响。在使用前测或多重测量时，测验可能成为影响内部效度的一个威胁因素。假设你对广告宣传是否会增加公司某种品牌的剃须膏的公众知名度感兴趣。你选取了一个大的消费者样本并发给他们一份问卷，问了一些关于不同品牌剃须膏以及与品牌有关的商业方面问题。通过为期3个月的一系列新的推销广告宣传后，你又一次将问卷发给同一批人，并发现他们更熟悉这种剃须膏产品。于是，你就宣布广告宣传是成功的。果真如此吗？

结论的一个问题是，广告宣传引起的品牌意识变化可能是因为前测本身导致了意识的变化。前测可能使得这批人更加敏感，从而更加注意这种剃须膏的牌子。在测试后的3个月中，他们可能更密切地关注所有关于剃须膏的商业广告，因此，在3个月后，他们能够说出更多关于每个品牌的信息，而不论之前有没有见过新的广告宣传。

测验除了使被试更加敏感外，它同样告诉了被试实验者感兴趣的主题，甚至实验假设。前测同样可以提供信息，并增加被试对于该主题的知识，因此，后测分数会很高，而这些变化与实验操作无关。

统计回归

统计回归可能是最为微妙的影响内部效度的因素。统计回归指的是当实验者根据人们在某一测试上的高分或低分选择被试，在第二次测试中被测者的分数有向平均数靠近的趋向。测验分数向平均数回归的原因还不是非常容易理解，下面通过一个例子来帮助我们说明其中的原因。

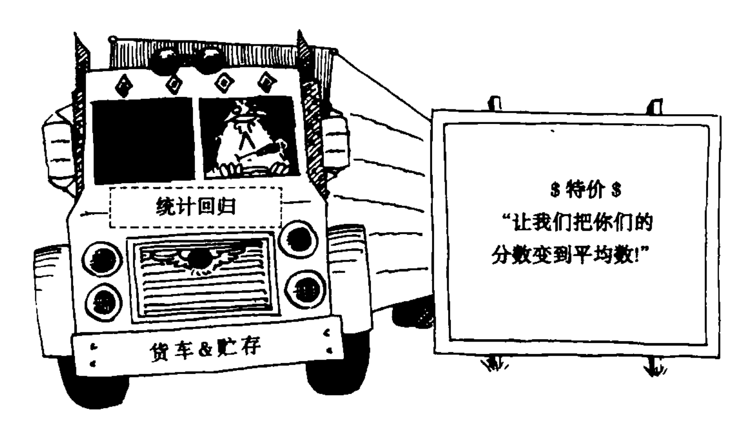
假设你已经制订了一个计划，该计划将会提高那些具有轻度智力障碍（智商为53～68）学前儿童的智商分数。你对这些儿童进行智力测验，并选出30名在轻微智障范围内的儿童。在1年以后，再对这些儿童进行智力测验。你发现这个组的平均智商会提高了7分，并且这种变化在统计上是显著的。于是，你就会宣布你的计划成功了。是这样吗？②

统计回归是如何导致或促成这一结果的呢？请你设想一下，前测的智力是由两个独立成分组成，即一个理想的测验可以测量出“真”智力和“误差”。在采用理想测验对某一特定儿童测试时，每次测验的分数应该相同。如果你使用这种理想测

---

① 除了我这里所强调的亡失率对内部效度的威胁外，你应该能够发现其他潜在威胁。例如，这个训练方案可能含有需求特征（见第4章），这些需求特质使得管理者在报告与压力有关的健康问题上出现偏差。

② 因此，你应该能够识别除统计回归外的其他影响内部效度的潜在因素。学前儿童在1年期间的成熟相当明显。测验同样是一个问题。智力前测很有可能是这些孩子已做过的一个测试。他们可能已经学了一些有关做测验的知识。他们也可能记住了前测中的某些具体题目，并且在1年内学会了答案。



验，就不会产生统计回归。但是，测得的智力分数也同样包括一个误差的成分。

这个误差可能是由一些不可预测的变量引起。例如，这个儿童可能比较幸运并且猜对前测中的很多条目；或者可能不幸运，猜中的几率小；或者那天上午测验者情绪特别不好，没有与儿童建立和谐融洽的关系。由于我们不能预测任何特定分数中误差成分的大小和方向①，我们必须把误差看成是从帽子里抽取的随机数字，并且在真分数上加上或减去它。

当你在前测低分的基础上选择一组智力轻度障碍儿童时，很有可能选择了那些不适当的儿童，而不是那些真实分数受人为因素影响的儿童。也就是说，该组的平均真分数高于他们的前测分数。因为你仅仅选了低分数儿童，你使组产生偏差。但是，在一年后的重测中，我们希望偏差少一些。我们期望使真分数增加的误差与使之减少的误差同样多。它仍然是错误成分，但是现在不再是使测量分数偏离真实分数的偏差。

如果还不确信，请你试着做一个小小的示范。拿出任何一个真分数，如 100。在相同大小的纸片上写下 -10 到 +10 这些数字，把它们放进一个容器。从容器中任意抽取一个数字，在 100 中加上或减去它，写下结果并代替原数字。在做了 30 次后，选出最小的 5 个数并计算它们的平均数(加起来除以 5)。现在以同样的程序选出 5 个数字计算平均数。第一个平均数小于第二个平均数吗？这样就可以证明统计回归。

与选择的交互作用

最后，威胁效度的因素，如成熟和历史等可能与选择产生交互影响。有关选择与历史因素可能交互作用的例子请看下面的研究。Coren 和他的同事们通过研究档案记录发现，在 10 岁到 80 岁的年龄组内左右利手的分布各不相同(Coren & Halpern, 1991; Porac & Coren, 1981)。10 岁年龄组有 15% 的左利手，而在 80 岁年龄组中降到了 0%。他们认为，左利手是"下降的生存适应"，这种适应导致他们在年轻的时候就死亡。很明显，许多左利手、左利手的母亲、左利手的丈夫都关注这一结论。但是，Lauren 和 Harris (1993a)对该结论提出质疑，并提供了相应证据，即历史

---

① 如果你用的是标准化测验，也许你能够知道这个测验误差成分的大概程度，即一个描述测验信度的数字。这个数字越小，我们越是应该考虑统计回归效应。

和选择的交互作用可能导致百分数的变化。80 年前，人们以左利手为耻辱。因此，家长和老师强烈鼓励儿童成为右利手，强迫他们用右手吃饭、写字和做其他事情。另一方面，右利手是社会压力选择的结果。但这一选择会随着历史而发生变化。多年以来，左利手变得更容易被接受，越来越少的家长和老师迫使左利手儿童成为右利手。所以 Harris 认为，到老年组几乎没有左利手不是因为大部分左利手死亡，而是因为一开始就很少。虽然争论仍在持续（Halpern & Coren，1993；Harris，1993b），但左利手消失的例子为我们提供了一个选择与历史可能存在交互作用的有趣例子。

我希望关于内部效度各种影响因素的讨论将帮助你确定混淆变量。当你设计实验时，检查每一个威胁因素可能有助于确定它们在实验中不会产生问题。有些情况下，也可能存在一些很难或不可能消除的潜在影响因素；此时，你可以采用准实验设计，这方面的内容将在第 10 章中讨论。

实验法小结

如果你已经熟悉实验法的用途，那么，让我们试着用框图描述已经学过的术语。图 2-1 总结了实验模式。图的左边列出了所有可能影响行为的条件，右边列出了所有潜在的行为测量。图的左上面选择其中一种条件作为自变量，右边是一个测量行为作为因变量。箭头表示我们对自变量是否导致因变量的变化感兴趣。虽然我们可以忽视其他行为，但我们必须确保能够解释其他条件。图中将这些条件划分为控

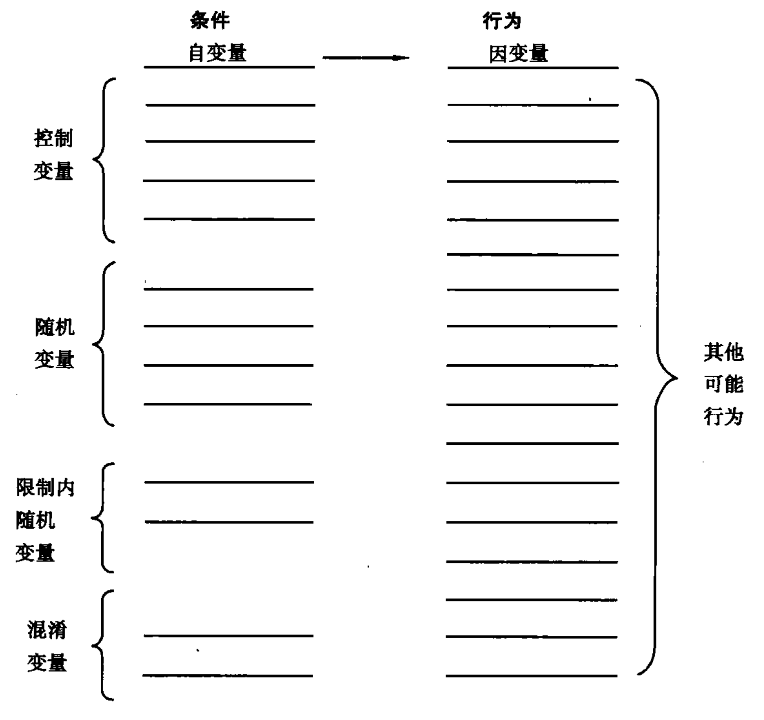


图 2-1　实验模式图。其中的一个变量被选做自变量。其他的变量被划分为控制变量、随机变量、限制内随机变量和混淆变量等。选择一种行为作为因变量。

制变量、随机变量、限制内随机变量和混淆变量。在划分变量的同时，我们应该记住控制增加了结果的精准度（内部效度），而降低了它的普遍性（外部效度）。在连续体的另一端，随机化降低了精准度而增加了普遍性。

最后，我以我和我的两名同事开展的一项实验(Grobe, Pettibone, & Martin, 1973)作一个说明实验变量类型的例子，并在一张类似于图 2-1 的图中列出一些变量。我们对教师讲课速度是否能够导致学生课堂注意力差异进行研究。那个时候我教一个由 200 名学生构成的班级的心理学导论课程，因此，我可以很方便地以不同的速度给这个班级讲课。我们选择 3 种不同的讲课速度作为自变量。在每一次课上，我试着以每种速度讲课至少 5 分钟。我们将每节课的这些部分录音，并且数出每分钟的音节数，以确保我的速度在允许的错误范围内。在图 2-2 中，将看到讲课速度列为自变量。我们可以用很多方法来测量学生的注意力，我们对学生进行录像，评判者能够通过录像推断出他们的注意力；学生们会完成一份反映其在每堂课上注意力的问卷，等等。因此，我们可能选择多种行为作为因变量。为了得到可靠的量化测量，我们记录了教室里背景噪音的水平，并推断当学生最安静的时候，他们注意力最为集中。因此，在图 2-2 中你会发现噪音水平作为因变量被列在行为之中。

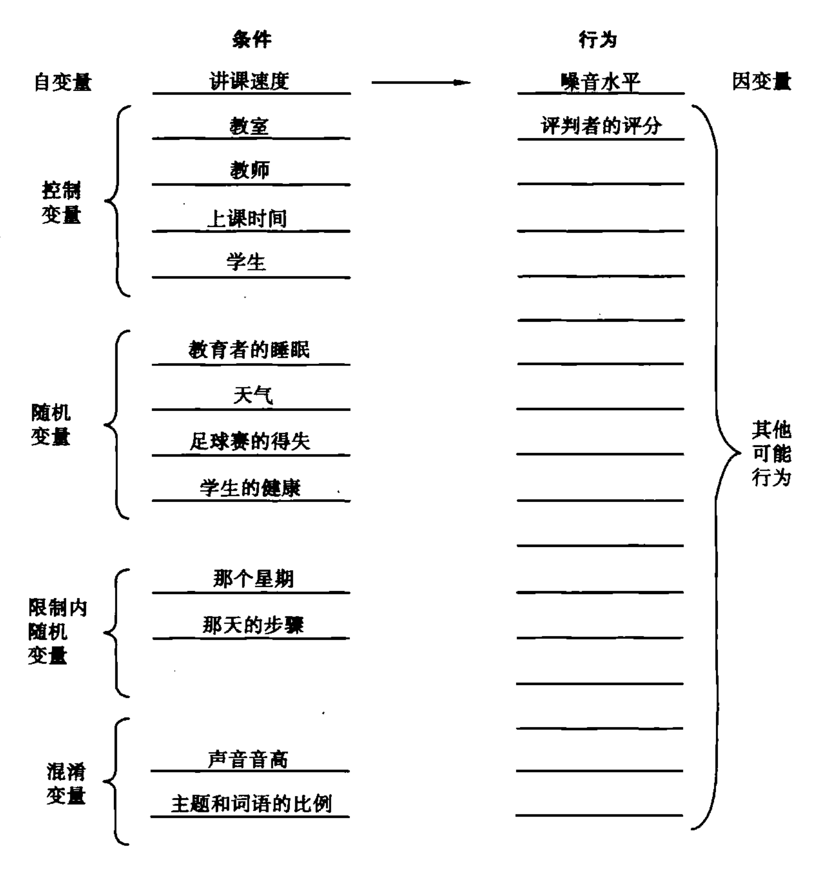


图 2-2 讲课速度实验中的变量包括一个自变量、控制变量、随机变量、限制内随机变量、混淆变量和一个因变量。

在实验过程中许多变量成为控制变量，但他们并未发生变化，如教室、教师、授

课时间段、教室里的学生等。这其中的一些被列在图 2-2 中作为控制变量。对其他有些变量我们没有进行控制,并希望它们以随机方式变化,如讲课前一天晚上我的睡眠情况、教室外面的天气、每个星期哪个球队胜利、班里面有多少人感冒(和大声咳嗽)以及许多其他因素。在图中有些变量被列为随机变量,其中包括一个限制随机变量。因为我们担心一个星期中的哪一天可能影响注意力,我们不想让所有慢速度讲课出现在星期一,中等速度在星期三,最快速的在星期五。因此,我们在限制范围内对每周运用每种速度讲课的时间进行了随机化,这样使得每天的每种速度有相同次数。

最后,虽然我们试图将混淆变量变到最小,但我们知道,与许多应用实验一样,有些混淆变量仍然存在。毋庸置疑的一个混淆变量就是我发音的平均音高。我不是一个机器,因此,与大部分人一样,我说得越快,声音变得就越高。我确信声音音高与讲课速度混淆。另外,只要课程的长度保持恒定,当我说得较快时,我可能要么说更多的关于特定话题的词语,要么说相同数量的词语并改变我讲解主题的数量。我试着去做到前者;所以,主题与词语的比率必然会与讲课速度相混淆。在图 2-2 中也同样列出了这两个混淆变量。我希望这个例子可以说明如何将变量划分为不同类型的条件和行为①。

小　结

实验法可以作出因果结论,即操纵一个条件会导致行为变化。操作的条件被称为**自变量**,它至少有两个水平。测量的行为叫做因变量,因为它可能依赖于不同自变量的水平而变化。自变量与因变量间的预测关系被称为假设。如果预测仅仅是自变量会引起因变量的变化,那么,它是一种无方向性假设;但如果预测是关于变化的方向,那么就是方向性假设。其他有些变量被叫做控制变量,它们被设置在一个特定水平,不允许其变化。还有些变量是随机变量,它们是随机变化的变量。一般来说,随机变量可以提高实验的外部效度,结论可以推广到其他人群、环境和时间。有些随机变量是在一定限制范围内的随机变量,它们可以在实验者设置的限制范围内随机变化。实验者应该设法消除或减少那些随自变量而发生系统变化、扭曲自变量与因变量间关系的混淆变量。

混淆变量会降低内部效度,使得我们无法确定因变量的变化是否仅仅由自变量引起的。内部效度的影响因素包括历史,实验期间发生不可控事件;成熟,实验期间个体的成长和经历;亡失率,组内个体非随机的丢失;测验,测验过程导致的被试变化;统计回归,在极端分数基础上选择组的分数趋向平均数;与选择的交互作用,不对等组的差别效应。

---

① 有读者想知道这个实验的结果。简单地说,我们发现,讲课速度确实影响注意力。幸运地是,中等速度时周围的噪声水平是最低的;速度最快时噪声水平也最高。因此,我们推断中等语速是最好的;与速度太快相比,还是速度慢些好。

3 如何形成一个实验构想

仅仅通过看,你就可以发现许多事情。

YOGI BERRA

力求完善是良好开端最大的敌人。

佚名

当他使用实验法检验第一个研究假设时,可能会有大量的其他假设涌上心头;而当他检验这些新假设时,又会有更多的假设浮现在脑海中;接着当他再次检验新假设时,又会有更多的假设浮现在脑海;直至最后,在不断地检验、排除和证实假设的过程中,他逐渐痛苦地发觉,其实假设本身的数量并没有减少。事实上,随着研究的深入,假设的数量反而增加了(increased)。

R. M. PIRSIG (1975)

知识的岛屿越大,其未知的海岸线便越长。

JOHN DONNE

我们对任何事情都知之甚少。

THOMAS ALVA EDISON

当爱因斯坦(Einstein)还是个孩子的时候,他便告诉我,他曾经设想过一个人追随着光线奔跑和一个人被关在下落电梯中的场景。其中,有关追随光线奔跑的构想导致了狭义相对论,而有关在下落电梯中的人的构想则导致了广义相对论。

L. INFELD (1950)

[福尔摩斯(Holmes)]:我现在还没有资料。在获得资料之前就提出设想是极其错误的。因为人们会略微调整事实以佐证自己的设想,而不是去调整设想以保持与事实的一致。

A. CONAN DOYLE (1891—1989)

作为一项学业要求，我曾经莽撞而愚蠢地在“心理学概论”课程中要求学生提出7个实验构想。起初，学生对这项作业的反应令我感到困惑。除了因情绪激动而咬牙切齿地聒噪、抱怨和叹息之外，我还听到那些被吓得目瞪口呆的学生发出了痛苦的哀号，“我怎样才能获得一个构想呢?”我不仅不理解为什么获取一个实验构想会成为一个大问题，而且我还发觉要解答这个问题几乎是不可能的。如今经过对这一普遍问题的深思之后，我已经对该问题发生的原因以及解决的方法形成了自己的观点。

我并不认为这一问题是因为学生没有想法而造成的。在孩提时代，我们对包括人类行为在内的每一件事情都会感到好奇：“妈妈，为什么那个人那么胖?”“Jenny用左手怎么吃饭啊?”“为什么我的拼写不如Betty那么好?”“为什么Tommy的父母老是打他屁股?”我不愿相信好奇心会随着年龄的增长而消退。事实上，那些觉得自己“无法获得实验构想”的学生在聚会或闲谈中有着相当多的想法：“应对生物考试的最好的学习方法是什么?”“我应该嫁给他还是只跟他同居呢?”“我是不是在清晨更有创造力呢?”

基于这一原因，如果你告诉我，你没有任何实验构想，我是不会相信的。你不是没有想法，而是可能害怕你的想法有错误！这种恐惧会抑制你天生的创造力，而不久之后，你所有的想法也会变得欠缺不当了。

害怕实验构想

对实验构想的恐惧往往是非理性的，是出于对心理学的一种误解。心理学家将非理性的恐惧称之为恐怖症(phobias)。由于我是一名心理学家，所以我禁不住诱惑，希望能对各种阻碍人们获得实验构想的恐怖症加以命名。以下列举的是最为常见的几种恐怖症①：

天才恐怖症(Geniephobia)(害怕天才)

天才恐怖症源于一种普遍的观念，即凡是做研究的人必定都是天才，因而这是中等智力水平的你所不能企及的。研究者通常都不会去抵制这种信念，反之，少数一些人还在培养人们的这种观念。多年以来，每当我阅读期刊文章的时候，我都会在头脑中将作者想象成一名白须飘然、长相睿智的老者。而当我发现许多实验者其实是一些年轻的、长相平凡的、像我们一样会做傻事说傻话的人时，我感到很惊讶。

现在我自己的天才恐怖症正在逐渐被治愈。所见的实验心理学家越多，我越不会相信只有天才才能做此类工作②。所以，放松吧！你的构想可能跟他们刚开始做这项工作时一样好。

① 如果这些名字与心理学界公认的专用术语相似的话，那么，此情况纯属巧合。

② 我并不是暗指实验心理学家比其他科学家更愚钝。生物学家和物理学家一样也可能是愚钝的。



在患上恐怖症之前，人会有许多想法

模仿恐怖症(Imitatophobia)(害怕模仿)

患有模仿恐怖症的人害怕提出任何观点，除非这个观点完全是原创性的。如果模仿恐怖症患者还担心任何值得做的实验都已经被别人想到了，那么，他们往往就会陷入完全瘫痪的状态。而事实上，在心理学中，真正原创性的实验确实是很少的。

大多数实验都是借鉴了别人的许多方法检验了另一些人的理论罢了。在第6章中，你将会了解到如何查找那些在你的兴趣领域中已经开展的实验，并且你还会发现你所从事的工作究竟有多么非独创。但是，不要害怕以小步幅向前推进科学。这就是我们这些人所做的事情。

设备恐怖症(Paraphernaliophobia)(害怕仪器)

动手恐怖症(Manuphobia)(害怕动手做事)

如果你现在所有的机械知识仅仅是汽车右边的踏板负责让车开动，而左边的踏板主管刹车，那么，你基本上可以算是一位准设备恐怖症患者了。一旦实验构想涉及任何比纸牌更复杂精密些的设施，那么，这种病症就会把你的构想吓跑。另一方面，如果不使用复杂精密的科学设备，你就不考虑做任何实验的话，那么，你就是另一种相反的疾病，即动手恐怖症的受害者。人们都知道，设备越复杂，研究就会做得越好。

但是这两种恐怖症都是毫无道理的。一些最棒的研究就很少或根本没有用到仪器。Jean Piaget 仅仅使用玩具木块、水杯和橡皮泥就在儿童心理学里开辟出一个卓越的研究领域。

其他一些心理学领域，如言语学习、概念形成、态度测评，以及人格等，其研究所需的工具无非就是笔和纸。仪器可以帮助你实施研究，但它并不是研究本身。同样的，如果在实验中必须使用仪器的话，也将会有人能对你加以指导。

简约恐怖症(Parsimoniophobia)(害怕简单)

简约恐怖症患者认为，他们必须以迅雷不及掩耳之势提出能够影响科学进程的宏大的实验构想。他们的信条是：如果简单，那就不是科学。尽管从事复杂的实验具有许多好处，然而，你通常还是应该瞄准那些能解答你的实验问题的最简单的实验。患有简约恐怖症的人很少能完成他们的宏大实验；即使完成了，他们往往也无

法对实验结果作出解释。那么，在开始的时候，就想得简单一点儿吧。你可以在以后任何时候继续去从事那些复杂问题的研究。（我将在第9章对简单实验和复杂实验作出解释。）



一个计算恐怖症患者

计算恐怖症(Calculatophobia)(害怕统计)

一些人害怕做任何比数手指更困难的计算。如果你从来都记不清如何计算汽车的加油里程，或不能将自己的支票簿整理得井然有序，那么，你就可能是一位潜在的计算恐怖症患者。如果你只考虑那些通过最简单的统计检验就能得出结果的实验，那么，请记住，这些测验只是协助你解释实验结果的工具而已；它们不应该使你放弃优秀的实验构想。你总可以找到喜欢玩数字游戏的人来帮助你分析数据。我并不是想说统计知识不重要，而是它毕竟只是一种科学工具罢了，统计并不是科学本身。

缺陷恐怖症(Imperfectaphobia)(害怕不完美)

如果没有将任何一个细节都处理得很完美，缺陷恐怖症患者就不会告诉你他（或她）的实验构想，而且他们的研究方案看起来就像是最终的实验报告。这种态度往往是源于他们阅读了太多以某种固定格式撰写的期刊文章。正如我们将在第5章所看到的那样，期刊文章是最终产品；它们很少会反映作者在实验之前所持有的草率想法及其通常的困惑。已完成的实验往往与研究者最初的想法具有极大不同。最初的实验构想仅仅成为了问题的核心内容；而当你提出实验以及实施实验的时候，实验的步骤也会跟着逐渐发展。如果你敢于冒险尝试，将自己的实验粗线条地描述给别人听，那么，别人可能会帮你将其改造为一个完美的实验，准确地说，是近乎完美的实验。

伪科学恐怖症(Pseudononphonosicentiaphobia)(害怕听起来不科学)

受这一可怕病症困扰的人们认为，只有以科学术语(scientific jargonese)①表述的想法才能称得上优秀的构想。科学术语是由科学家所编撰的、与同行交流时听起来比较酷的一种虚拟语言。它的作用是使研究变得晦涩、脱离大众——有时候也会使其他一些科学家感到费解。例如，对于一项评判“当单词以成组的形式出现时，人们的记忆效果是否更好”的实验，如果以术语加以说明便是，一项有关“分类与类别群集对言语材料保持的影响”的调查。或是一个有关“迫于朋友的压力，少数种族的人会居住在同一社区”的概念，可以被描述为，一项检验“种族联系所具有的同伴群体强制力功能对人口统计学分布造成的影响”的实验。专业术语很容易翻译

① 我用术语(jargonese)一词指代字典中专有名词(jargon)一词的含义，意思是指“以独特性、复杂的学名和有关语法为特点的口头或文字作品，”而不是指“在特殊的贸易、职业或组织中使用的特有语言。”术语与专有名词的含义确实差别不大。

成日常用语。如果你对"从属偏爱影响人们感知觉维度突出性"的问题感兴趣,那么,你实际上是在查明"加入特定组织中的人是否会以不同的视角看待别人"。试着自己翻译下面这句话:"母亲工作对同胞兄弟姐妹的攻击性趋势的影响。"①

活动恐怖症(Ergophobia)(害怕工作)

很遗憾,对于这种病症,没有已知的治疗方式。

现在我们已经认识到哪些恐惧可能会妨碍我们的创造力,那么,让我们试着努力获得一些实验构想吧。在开始时,最好使用怎样的方法呢?

观　察

有人曾经说,写作是件容易的事情:只需要坐在电脑前,盯着键盘,一直等到血滴(drop of blood)从前额上渗出。这句话也同样是描绘避免实验构想产生的最好方法。由于我们对人类的行为比对键盘的行为更感兴趣,所以,我们最应当做的事情是去观察人类而不是去观察键盘!

产生实验构想只不过是去注意到你周围发生的事情。一旦你成为了一名优秀的观察者,那么,你天生的好奇心就会为你提供可验证的实验问题。为时一周的持续观察应该能为你提供足够的实验设想,让你一辈子都做不完。

事实上,实验心理学中一些经典研究都始于简单的观察。如果 Eckhard Hess 的妻子没有注意到,当他观看鸟的图片时瞳孔会变大,那么,瞳孔计量仪就可能永远都不会被发明。如果 Ivan Pavlov 没有注意到,当他的狗在见到肉沫之外的刺激物时会分泌过量的唾液,那么,Igor Nosnoranovitch 就可能会取而代之,成为经典条件作用的创始人。如果 Jean Piaget 没有注意到,当他女儿 Jacqueline 看不见奶瓶时就会停止发出咯咯的笑声,那么,他可能就不会成为瑞士著名的观察巨匠。大多数革新性的实验构想均来自于简单的观察。

公共场所的观察

在阅读完下面几个段落以后,请带上纸和笔,离开你的房间,到外面有人可以观察的地方去走走。作为一项观察训练,请记下你散步时所想到的任何实验问题。

首先,我会出去闲逛一下并借此向你说明我的用意:我在外面散步,看到阳光普照大地。

1. 当天气好的时候,人们完成的工作是更多还是更少?

我从 2 名正在为自行车道铺设混凝土的工人身边经过。其中一位正在工作,而另外一位则站在旁边看着。

2. 是不是和别人一起工作的时候,工人会更多地站在一旁呢?

有两个慢跑的人从我身边经过。

---

① 如果你的回答与"是不是母亲进入职场,孩子就会发生更多争执?"相似的话,那么,你就已经理解该问题了。记得一定购买我的下一本书:《为乐趣与益处所使用的科学术语》(*Scientific Jargonese for fun and profit*)。

3. 经常做运动的人是不是晚上睡得更香呢?

一位年轻的女士和一位蓄着胡子的年轻人坐在那边的树下。他们看起来非常恩爱。我觉得自己有点像"偷窥的 TOM",所以,我决定最好还是继续往前走。

4. 女人是不是觉得留胡子的男人比没有留胡子的男人更有魅力?

我看到一大群学生涌入教室。

5. 在大型班级上课的学生是否会比在小型班级上课的学生成绩更好?

我来到人行横道前。那辆车要停下吗? 是的,它停了。于是我穿过了马路。

6. 司机是否会更多给异性行人让路呢?

我停下来,看着一辆跑车嗡嗡地飞速驶过街道。

7. 驾驶跑车的人是否比驾驶普通车的人开车速度更快?

我往回走,经过了图书馆。

8. 与在宿舍学习相比,在图书馆学习的学生的记忆效果是否更好呢?

我从办公楼前的自行车架旁走过,看到了许多辆自行车。

9. 山地自行车是否比公路自行车更好骑呢?

我大步跑上楼,进入自己的办公室。我回来了。

我刚刚获得了 9 个潜在的实验构想。几乎是每分钟一个! 现在我等待着你去尝试。

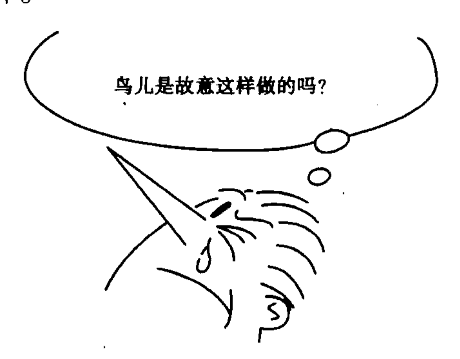

我在等待

欢迎回来。你有很多想法了吗? 设想一下,如果你一直都这样观察的话,你将获得多少构想。然而,不是所有的重要问题都能通过实验来解答,所以,现在所面临的问题是"我应该将哪个想法转化为实验呢?"所有的实验问题都必须通过 ROT 检验,即它们必须是可重复的(repeatable)、可观测的(Observable)和可验证的(Testable)。有些问题之所以不能成为实验问题是因为它们不具备可重复性。例如,一些超感官知觉(ESP)的拥护者宣称,这种知觉只在特定的条件下产生,并且条件适当的时机几乎无法预测。换言之,ESP 只在某些时候发挥作用,因而它没有通过可重复性检验。只要这种基本的信条控制着 ESP 的结果,那么,就不可能检验这一能力是否存在。另一些问题无法成为实验问题,因为它们不具备可观测性:"狗是否能像人一样思考呢?""我对红色的体验与你的一样吗?"最后,一些问题没有成为实验问题,是因为它们不具备可验证性。例如,科学无法解答道德问题,如"堕胎是错误的吗?""女人穿短裙子合适吗?""毒品是邪恶的吗?"尽管我们的确能用科学的方法判定人们对这些问题的看法;然而,我们却不能设计出任何测验对这些问题本身作出解答。因此,我们必须从实验构想清单上排除所有这样的问题。

你的想法清单中的所有问题是否都达到了 ROT 要求呢? 花点儿时间仔细检查

你的清单，并把那些没有达到要求的问题淘汰掉。

读完第 1 章后，你也应该意识到，一些问题必须通过相关性的观测而不是通过实验来解答。例如，倘若我们要回答第 7 个问题：选择驾驶跑车的人是否比驾驶其他类型汽车的人开车速度更快，那么，我们就必须进行相关性的观测。另一方面，如果我们想回答下面这个的问题：人们在驾驶跑车时是否会比驾驶其他车时将车开得更快，我们就可以设计一个实验。再看一下你的想法清单，给每个想法都做上实验性或相关性的标记。

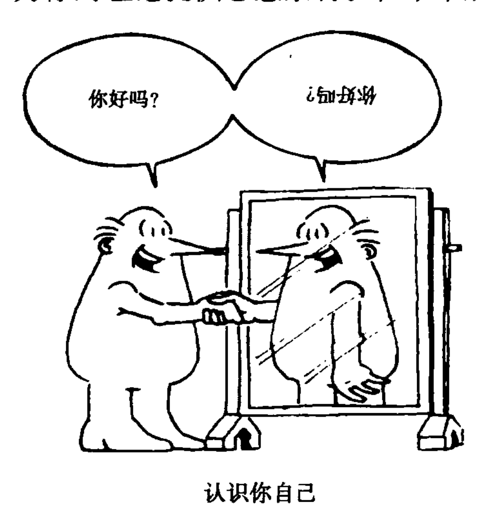


并不是所有的问题都可以造就好的实验

我们的一个小小散步非常有趣，不过人们在公共场合所展现的行为毕竟有限，那么，我们还能去观察谁呢？

观察自己

内省法是实验心理中最早应用的技术之一。内省者专注于观察自己的心理过程，而不是观察自己的行为。在一场有关"人们是否能够认知自己的心理过程"的争论之后，实验心理学家完全停止了自我观察。他们不再信奉"认识你自己"的格言，而是断然坚持"不认识你自己"的信条①。作为一名心理学家，倘若你想成为自己实验中的唯一被试，这将无法得到普遍的认同；然而，你却可以通过该方式获取很多优秀的实验构想。你不仅能够由此而收集到许多你感兴趣的行为样本，而且你还可能会认识到自己为何作出这样的行为。前者可以给你提供实验构想，而后者则可以为你的理论提供思想素材。在本章后半部分，我们将对理论进行分析。



认识你自己

稍微努力一下，你就可以开始观察自己的行为了。虽然说你没有注意过自己，这听起来似乎有些荒谬，但这很可能是真的。当你穿衣服的时候，你先把哪只胳膊伸进衬衫呢？当你刷牙的时候，你先刷左边还是右边呢？当你拿房间钥匙开锁的时候，是朝逆时针方向转还是朝顺时针方向转？当你交叉双腿的时候，是常把左腿放在上面还是把右腿放在上面？这些都是你每天所做的事情。你注意过它们吗？观察自己是很有趣的②，同时它也是实验构想的良好来源。当你产生想法的时候，

① 一些实验心理学家仍然不知道他们是谁。

② 如果你具备了这种技能，你需要在公共场合下控制自己的行为。倘若你因自己的行为而阵阵大笑，会被别人认为是不正常的。

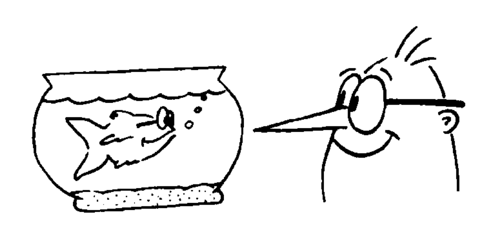
就将其记录下来。

观察朋友

你的朋友也是获得实验构想的良好来源。然而，重要的是应当尽可能不唐突地去观察他们的举动。盯着别人看，在最好的情况下会被认为是不礼貌的，而在最坏的情况下则可能会引起纷争。有时候人们会不太喜欢自己的行为太引人注目，所以，他们可能不想被别人观察。因此，为了避免失去朋友，还是悄悄地观察。不要过于显示自己的洞察力，因为无论它有多么杰出，它都无法帮你赢得朋友或影响他人。

观察儿童

如果你有兴趣在发展心理学领域开展实验，那么，就有必要去观察儿童。不过儿童也可以为其他的研究领域提供良好的构想。如果你不幸没有孩子的话①，那么，你的亲戚和朋友可能会非常乐意让你去观察一段时间他们的孩子。儿童的行为方式与成人不同，成人知道在外部观察者面前，自己的行为应该表现得理性、合乎逻辑、前后一致，而儿童的行为方式通常比较简单，并且没有受到复杂模式或社会禁忌的影响。由于大多数幼儿都不关心成人的标准，所以，你可以在儿童身上观察到相对没有受到污染的行为模式。



观察你的宠物

观察宠物

出于其自身的特性，动物可以成为有趣的研究对象，而且它们的许多行为也可以推广到人类。另外，你还会发现宠物比儿童更加无拘无束。因为它们不具备高度复杂的行为模式，所以，它们的行为通常更容易解释。此外，你可以操控宠物的环境，而无须担心可能存在的永久性伤害所引发的道德问题。（见第 4 章有关"对待动物的道德规范"的探讨）

替代观察

虽然阅读别人的研究没有直接观察那么令人兴奋，不过你还是可以通过它获得实验构想。你也许会觉得替代观察（vicarious observation）这种技术是在依赖别人的创造力，但尽管如此，该方法具有一定的实际操作优势。首先，由于你所提出的主要实验问题已经有了作者和期刊评审人认可的迹象，所以，你能够了解到自己提出的问题是重要的。其次，别人已经将实验结果纳入到现有知识体系之中，并且可能提出了某项理论，因此，你的研究领域已经形成了体系，这样可以节省你的时间与精力。最后，早期研究者已经开发了测量攻击行为的方法，并且该方法显然是较为有效的，或许你可以将其修改后运用于自己的研究中。

当你寻找实验构想的时候，你应该首先确定自己感兴趣的研究领域。然后，你

① 或者你没有受到孩子的烦扰（这取决于你的观点）。

会了解到自己应该阅读何种类型的期刊。你的研究主题应当尽可能地明确:小组内的竞争、学校中的游戏疗法、视错觉的感知以及算术能力的发展等。对一些更概括的主题而言,你可以简单地浏览与之相关的期刊。而对于更明确的主题来说,这一程序却不那么奏效,这时你需要进行如本书第6章所描述的那种文献检索。

当你阅读这些文献的时候,尽力去挖掘那些研究中没有解决的重要问题。有时作者也会通过未来研究取向的提议,协助你发现这些问题。不过在读完文章之后,你应该也能通过未解答的问题来确定研究的走向。通常情况下,你并不希望只是简单地重复已有的研究,然而,从事一些与之相似的研究还是无可厚非的。

通过例子来加以说明,假设你阅读了一篇有关"电视上的暴力节目会引发儿童攻击行为"的实验报告。在实验中,6岁儿童持续一个月每天接触两个小时的暴力电视节目或半小时的非暴力电视节目。通过他们所玩的玩具观测其行为,其中攻击性玩具包括枪支、刀子、坦克等,而非攻击性玩具包括娃娃、卡车和积木等。读过这篇文章,你也许会决定由自己提出一种更好的方式来操控自变量。或许你不喜欢他们定义暴力的方式,不喜欢他们选择给两组观众呈现的电视节目,而你可以以不同的方式来界定这些问题。或者你可能希望增设第三个实验组,让他们观看暴力与非暴力相结合的节目,或观看一套更加中性的节目,或者根本就不让他们看电视。

你也许想要改变因变量,而不是改变自变量。你可能觉得通过儿童所玩的玩具来确定其攻击水平并不是测量攻击的良好方法。或许你认为,如果由受过训练的评判者来观察共同游戏的儿童,并由他们评价每个儿童的攻击水平更为合适。或者你还希望去访谈儿童的老师或家长等。

你也许还会想到,一些控制变量并没有被调整到合适的水平。6岁儿童已经接触过很多电视节目。你可能认为更好是选取年幼的儿童或使用不同年龄的几组儿童来进行实验;或许你认为,对于平均每天看电视接近4小时的儿童来说,只看2小时的电视节目,时间太少;抑或你会认为,1个月的时间不足以显示电视对行为的影响效果。

你可能会想,应当对其中一个控制变量进行随机化处理。例如,你认为,研究者让儿童以6个人为一组在实验室里看电视,而不是让他们在处于家庭背景中的自己家观看电视是错误的。

如果你认为自己发现了一个能够排除的混淆变量,那么,你的研究就会变得更加有趣。例如,导致儿童攻击行为的原因可能仅仅在于暴力节目比非暴力节目的声音更大;或者是喧闹的噪音而非暴力行为致使儿童变得更具攻击性。你可以想出许多方法修改最初的实验,从而以不同的方式检验假设,详细地阐明假设或证明一个与原来的假设相似却不完全相同的假设。

因而,当你阅读他人的研究报告时或许会发现,仔细地审查与以下问题类似的疑问是颇有裨益的:在操控自变量、测量因变量、选择控制变量的水平、将控制变量转换为随机变量或反之、或在避免混淆变量等方面,是否存在其他更好的或不同的方式呢?换言之,你能够想到一些改善研究内外部效度的方法吗?我相信,如果你仔细阅读文献并询问自己这些问题,你就会产生许多出色的研究构想。

拓展你自己的研究

一旦你做过几个实验,你就会发现,自己的研究能够提供许多实验构想。你做的每一个实验都会留下一些未解决的问题。例如,当你在一项实验中处理过自变量的几个水平之后,你可能就希望了解使用自变量的其他水平会产生怎样的结果。或者你可能在一项实验中将某一变量控制在了一个特定的水平,那么,如果将其设置在不同水平将会发生什么。也许你的实验得到了意想不到的结果,你希望弄明白为何结果与预期不一致。一般来说,每个实验所带来的未解问题总是多于已解决的问题。

科学就是不断产生新问题,而不仅仅是解答固有知识体系中的问题。在后者的观念中,随着科学研究的开展,科学领域中的未解问题将会变得越来越少。然而,事实上,每项实验都增加了需要解答的问题数量。所以,我们需要应对的事情并没有减少;恰恰相反,我们被卷入了自己可能都无法掌控的更多的事务之中。

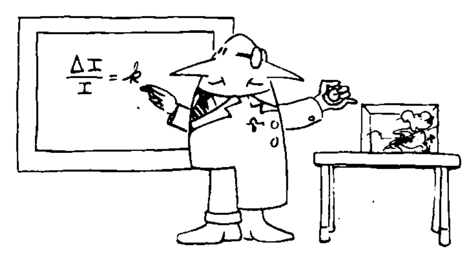
这种对科学的开放式观念既令人沮丧又令人兴奋。说它令人沮丧,是因为在这样一个不断扩大的科学领域中,有时我们所做的工作看起来似乎是后退五步才能前进一步,因而我们所取得的进展有时是难以描绘的。另一方面,说它令人兴奋,是因为我们最终在不停地探询一个比一个更好的问题。也许,科学的目的不在于解答所有可能的实验问题,而在于解答那些更有前景以及更为重要的问题。继续从事自己的研究,你就会发现,你所面临的主要问题不是"我如何才能获得一个实验构想?"而是"哪个实验构想是最重要的?"

使用理论获取构想

如果你已经收集了许多观察结果,那么,你如何将这些结果纳入到一个决定实验类型选择的理论框架中呢?在科学领域中,常见做法是提出一个理论。一般情况下,实验的目的便在于验证某一理论。所以,获取实验构想的一种方式就是将你的某些观察转化为理论;然后,再以一项实验来验证它。然而,令人遗憾的是,那些不了解理论是如何在科学领域中应用的人往往对这一方式持有消极看法。一种误解是,理论是极其复杂的,只有天才才能明白:"爱因斯坦可能会知道 $e = mc^2$ 代表什么意思,但我永远不会理解。"第二种误解是,理论只不过是某些人大致的猜想罢了:"那只是一种推测(theory)。"事实上,理论很容易被人理解,而且随着支持某个理论证据的不断累积,我们更加确信其真实性,但永远无法全然地肯定某一理论。

我们为什么需要理论?由实验和其他类型的研究所得到的结果都是事实。而科学并非随意地收集事实,它是有组织的知识;就像建筑物一样有一个结构。正如随意堆砌的砖块不是建筑一样,没有系统地收集事实证据也不是科学。理论提供了一份蓝图,它能告诉我们如何将这些事实归整到一个有组织的科学知识体系中。在我看来,实验心理学比其他一些科学更为有趣的一个原因是,实验心理学家不单纯是建筑师,还是施工人员和砖头制造者。而其他一些科学则具有不同的分工。例如,多数物理学家不是理论物理学家就是实验物理学家,却不是双项全能的;而实验心理学家一直以来就从事理论和实验两方面的工作。

要给理论下一个简单易懂的定义是比较困难的。如果必须这样做的话，那么，理论是对一组抽象变量之间可能存在关系的陈述。理论性的陈述具有或然性（probable）。这是因为它依旧受制于检验，理论更容易通过检验被证伪而非被证实。变量间的关系存在于抽象（abstract）变量之中，这是因为如果变量都是由直接可观测的事件构成的，那么，我们得到的只是有关事实的陈述、一种直接的观测而非是理论。理论中的抽象变量是关于环境或行为的一般范畴，而不是某些具体的环境或行为；正如理论性的陈述“观看暴力场景会导致攻击行为”与一项实验范例“在观看战争题材的电影之后，儿童会在游戏时选择使用枪械”之间的关系。



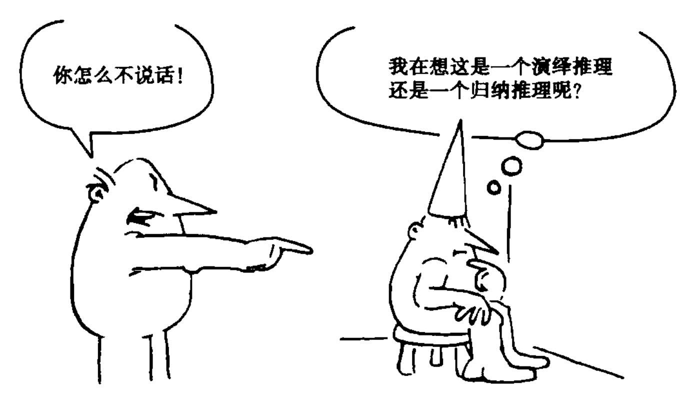
实验心理学家从事两项工作

为了帮助你更好地了解理论的应用，我将通过一个案例来说明如何建构理论以适用于实验的设计与解释。假设我们观察了周围的世界，并注意到以下事件：在观看了电视上枪战镜头比较多的影片之后，儿童的游戏行为变得更加暴力了。近年来，年幼儿童被指控实施暴力犯罪的频率越来越高，与此同时，媒体中的暴力行为似乎也有所上升。在饱受战乱的国家中，儿童在很小的年纪就拿起武器开始搏杀。这些观察结果——当然可能还有更多——或许会促使我们提出这样一种理论：“儿童观看到的暴力行为越多，就越有可能出现攻击行为”。

这一理论非常易于理解。但需要注意的是，它比任何一项对儿童行为的观察都更为抽象、更为概括。或许我们并没有意识到，我们是通过归纳法得出这一理论。归纳法（induction）是一种逻辑加工，通过该方法能够得出的结论比观察所涵盖的信息更多。也就是说，我们不仅可以预期“有关暴力与攻击行为”的理论能够被上述三项观察结果所支持，同时，也能被其他所有儿童观看暴力行为的案例所支持。当然，我们的预想可能是错误的。或许这一理论只适用于所观察的案例，而对我们所归纳的案例并不适用。然而，一旦建构了理论，至少可以在将来通过实验加以检验。

如果我们的理论行得通，它将允许我们提出许多假设。提出这些假设的逻辑加工被称为演绎法。当我们使用演绎法（deduction）时，我们可以从一组前提中得出一个结论，而该结论所包含的信息存在于所有前提之中。因而，如果前提中的信息是正确的话，那么，结论也一定是正确的。比如说，倘若马是哺乳动物，并且所有哺乳动物都是动物的话，那么，通过演绎推理可知，马一定是动物。在我们的例子中，如果儿童观看的暴力行为越多，他们就会变得越有攻击性，并且如果电视播放的刑侦节目中存在暴力行为，那么，大量观看电视中的刑侦节目则必然会导致攻击行为的增加。通过演绎法，我们可以从理论中推断出跟这一构想接近的众多观测行为。每个可预期的观测都可以形成一个实验假设（hypothesis）。为了检验假设，我们可以设计这样一个实验，让一组儿童每天观看 4 小时包含暴力行为的电视刑侦节目；让另一组儿童每天观看 4 小时不包含暴力内容的电视节目。在几天之后，观察儿童的游戏行为以测定他们的攻击性强度。由我们的理论所演绎得出的假设是，观看刑侦节目的实验组会表现出更多的攻击性行为。如果理论正确，那么，假设也必然正确。表 3-1 显示了到目前为止我们所提到的思维加工过程。至此，我们使用归纳法将观

察转化为理论,也使用了演绎法从理论推测出可预期的观察结果。现在我们就可以用实验来检验这一预测。



如果实验证实了预测的观察结果,那么,能否说理论得到验证了呢?还不行!一个假设的证实并不能说明生成该假设的理论得到了验证。假设的确支持了理论,不过只有通过归纳法——而非通过演绎法——我们才能得出这样的结论。要最终证实一个理论,就需要检验从理论演绎出的每一条假设。在刚才提到的案例中,我们必须要检验儿童观看暴力行为的每一种可能的方式,去测量儿童可能表现出的每一种类型的攻击行为。如果做不到这一点,我们只能说自己的实验结果为有关理论提供了支持。随着其他一些能为理论提供依据的实验的完成——尤其是其中那些检验多种变量实验的完成——我们对理论的信心也将继续增加,不过你仍然可以看出,要说某项理论得到了证实是多么困难。

另一方面,假如实验结果与预期的观察结果相悖,那么,我是否就能证明理论是错误的呢?按逻辑来说,我们确实证明了(再参见图 3-1)。只要有一个预测不成立,就可以对至少一个前提予以反驳,这是因为预测是通过演绎法得到的。还记得那个"马是哺乳动物,所有哺乳动物都是动物,那么,马一定是动物"的例子吗?如果我们发现,马不是动物,则要么并非所有哺乳动物都是动物,要么马不是哺乳动物,不会再有其他的逻辑可能了。按照相同的逻辑,与观看非暴力节目的儿童相比,那些观看 4 小时暴力节目的儿童并没有表现出更多的攻击行为则说明,要么我们给儿童呈现的暴力节目并不暴力,要么我们的理论表述是错误的。因此,为什么著名的科学哲学家 Karl Popper 认为,科学家的工作不在于证实理论,而在于证伪理论(Popper, 1968)。Popper 后来将这一观念简化为一条口号,"无论看到多少只白天鹅均不能证明'天鹅都是白色的'理论。哪怕仅看到一只黑天鹅也会将其推翻。"

通过实验从逻辑上反驳某项理论的主张的缺陷就在于,进行实验假设检验所使用的统计计算并不具有演绎性。尽管我将在第 12 章详细讨论这个话题,不过我仍要在此说明一些基本的论点。在我们提到的例子中,反驳实验假设的方法在于发现两组观看不同节目的儿童的攻击行为之间不存在差异,换言之,两组儿童的攻击行为在统计上不存在显著差异。但问题是我们的统计检验通常是用来发现自变量的不同水平是否会导致行为的差异(difference),而不是一致性①。但通过实验反驳假

① 从技术上说,当检验虚无假设的时候,一开始我们会假定变量之间没有差异,而统计检验将告诉我们,这一假设在多大概率上是错误的。然而,检验并没有告诉我们变量相等的概率有多大。

设，一般是要展现行为水平之间彼此相等而不是有差异的，而我们使用统计所检验的并非一致性。因此，基于这样的结果来反驳一个假设，继而反驳一个理论时一定要小心。

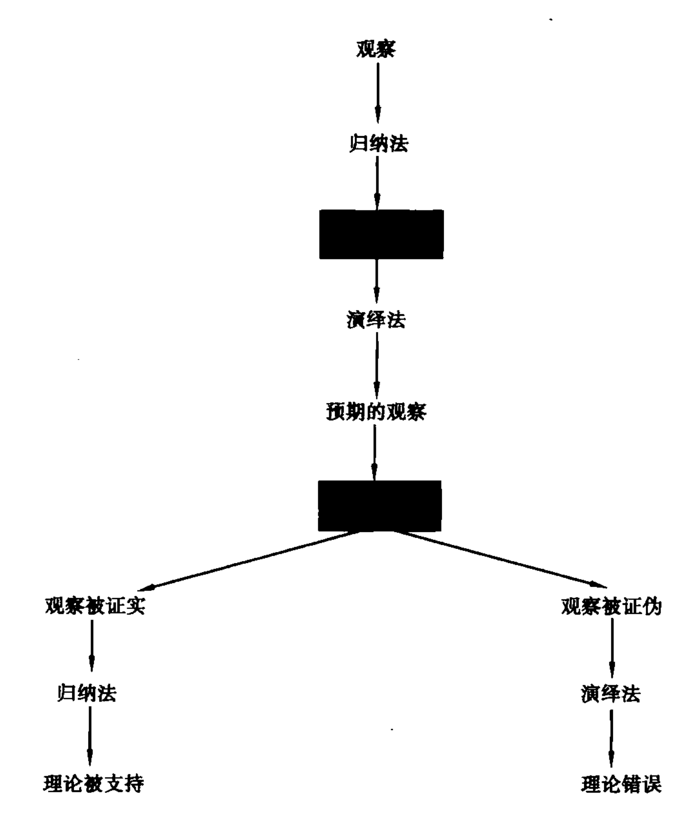


图 3-1 归纳法与演绎法在连接理论与实验中的作用

即使我们采取合适的统计路线（即根据统计显著性来否定一项假设）否定假设，我们仍然可能会犯错。就像我们将在第 12 章所探讨的那样，即便正确地使用了统计检验，其结果也只能在概率水平上正确。基于案例中出现的差异，我们只能在一定置信水平（certain level of confidence）上说，人群中的确存在着差异。例如，我们可以说，结果在 0.05 水平上具有统计显著性。因而，在 20 次反驳假设的过程中，我们可能有 1 次是错误的，并且我们认为理论被推翻的想法也可能是错误的。

除了上述的统计学原因之外，还有其他一些原因能够解释"为什么证明由演绎逻辑所推出的假设是错误的，并不能成为强有力的理论反证"。例如，在实验中对理论的考查方式，对暴力行为的操作化或对攻击行为的测量方式、或对变量的控制或随机化处理中都可能存在问题。因而，仅仅通过一项实验并不足以否定一个理论，只有收集到大量的反证才可以做到。

下面我将进一步讨论理论在帮助我们获取实验构想以及解释结果方面的作用。一般来说，了解理论和实验之间的相互影响是非常重要的。幸运的是，你不需要在每次实验之前都通过正式的逻辑加工去思考这个问题。对于任何实验而言，逻辑加工的步骤都是一样的。事实上，人们都会非常自然地经历我所描述的这个加工过程。为了生活，我们会不断地进行观察，会对观察的结果加以概括化，会通过更多的

观察去检验所概括出的结论。尽管我们没有把这些概括化的结论称之为理论,但我们的生活确实是建立在这些概括化的结论之上的。我在这里只不过是以一种更为正式的形式描绘了这些自然加工过程而已。

理论的类型

到目前为止,我仅仅给出了一个理论范例。而理论可以具有许多形式。我将在这里探讨三种类型的理论①,并继续以"电视暴力是否会导致攻击行为"这个命题为例来说明这些理论。

描述型理论

顾名思义,描述型理论(descriptive theory)是对事件加以命名,但没有对事件发生的原因或过程提供必然解释。例如,作为精神分析理论的一部分,Freud 曾经指出,当我们无意识地将那些引起痛苦或令人厌恶的想法排除于意识之外时,就会产生压抑。尽管这样的理论对临床医生有所帮助,然而,仅仅通过命名并不能解释导致压抑发生的条件,而且无法说明可以通过怎样的实验来对其加以考查。同样的,多年来,动机研究领域的心理学家一直痴迷于对各种本能加以命名。在开始的时候,命名这些本能似乎是有好处的,因为多数动物的行为都可以被归结为某些本能的反映(如喂食本能或择偶本能)。然而,随着命名的增多,一些心理学家最终也将一些可观测的行为(如"当迎面遭受攻击时,动物会逃入洞中"的本能)归结为本能,这便使此概念丧失了效用。



一种描述型理论

当描述型理论命名的是一些抽象性事件而非一些直接可观测事件时,它会显得更加有用。例如,我们可以说,观测到的暴力行为与攻击行为之间具有相关性。而如果我们能认真地将"暴力"和"攻击行为"定义为一系列概括化的事件,那么,我们或许就可以得到一个描述型理论。但即使是这样的描述型理论,其价值也非常有限,因为它并不能解释两者之间的关系是如何运作的。

类比型理论

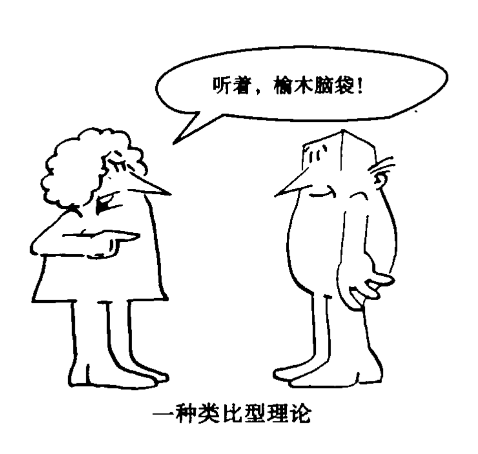
类比型理论(analogical theories)通过在心理关系与物理模型之间建立一种类比,解释两者间的关系是如何运作的,从而使物理类比变成有关行为的心理模型。例如,许多理论试图以计算机作为物理类比物,以说明人类加工信息的方式。当然,没有人真的相信,大脑会完全像计算机那样运作;然而,由于两者之间存在诸多相似之处,这使计算机模型为我们提供了一些有用的类比理论。

下面是一个类比型理论的例子。让我们以动力这种物理属性作为类比来说明暴力与攻击行为之间的关系。你可能知道,一个物体的动力大小与其速度和质量有

---

① 虽然一些名称会有所变化,但这里所探讨的三种理论类型与 Arnoult (1972)在其《心理学基础科学方法》一书中提到的理论类型比较接近。

关;其运动速度越快、质量越大,则它的动力便越大。而摩擦力可以克服动力。由此,我们可以将"暴力与攻击行为之间的关系"演化为这样一则类比型理论:"观察者所了解到的攻击行为的数量如同移动物体所生成的力,而观察到的暴力程度可以类比为物体的质量,观察到的暴力时间可以类比为物体的速度。在接触了一段时间的暴力行为之后,攻击性发展趋势可能会增高。然而,随着时间的推移,攻击行为又会减少,这就如同摩擦力会克服动力一样。"



一种类比型理论

这样一个类比型理论比前面所提到的描述型理论更为有用,因为它能解释一些复杂的关系。而且我们还可以在了解物理模型运作的基础上来检验这些理论。例如,我们知道,在一个物体上施加力的时间越长,则其运动速度越快,其动力越大,并且用来克服动力所需的摩擦时间也要越长。由此可以类比,人们观察暴力行为的时间越长,就需要越长的时间来消除攻击行为。

鉴于其解释效力,一个类比型理论自然要比一个描述型理论更为有用。但类比型理论仍然存在局限性,因为在某些时候,物理类比的特性可能不再与人类的特性相吻合。因此,你最好是将类比型理论作为初始的理论,以便鉴别主要研究变量,并描绘出变量之间相互影响的一般方式。不过,类比型理论很少能有效地帮助你确定变量之间精确的数学关系。

量化型理论

量化型理论(quantitative theories)尝试以数学术语来阐述事物之间的关系。它们不仅会详细地说明不同类型的变量之间的关系方向,而且还会说明这些不同类的变量是怎样在数量上相互联系的。很少有心理学理论能够达到这一复杂层次。只有在学习、记忆和认知等领域中才有一些心理学家在尝试着使用量化型理论。

量化型理论在心理学领域的使用有限,因为心理学家比物理学家更难应对变化性。例如,在物理学中,有一条用精确的数学术语界定的关于重力的量化型理论。由于所有物体受重力影响的方式相同,所以,物理学家可以假定实验结果的任何变化都是由测量误差所导致。然而,在心理学领域中,我们不能单纯依据某个人的行为来预测所有个体的行为,也不能因此而预测该个体在任何特定时间的行为。因而,我们的量化型理论必须能够适应这种变化性。我们能做的最好的工作是预测行为发生的概率①,因此,我们必须以概率术语来表达数学关系。(例如,某人在五次试验中学会这组单词的概率为0.8。)

心理学家还面临着挑选行为测量尺度的问题。在物理科学领域中,有关速度和

① 在一些领域中,物理科学家也面临着相似的问题。例如,通过概率来阐述原子的结构。事实上,混沌理论(一种基于非线性数学所建构的动态系统理论)的发展可以用来解决物理科学中存在的那些没有遵循经典科学规则的或然性的事件。

质量的测量单位并不存在争议。但在心理学领域中，我们必须要找到能够测量暴力或攻击行为等概念的尺度。例如，考虑下述量化型理论：人们所展现出的攻击性水平与其近期观看暴力行为的平均水平直接相关。由于我们试图通过理论在暴力行为与攻击行为指标之间建立一种数学关系，因而就必须先确定如何去测量它们。众所周知，为这样的概念确立测量尺度并不是一件容易的事情。

近年来，量化型理论在心理学的一些领域中得到迅速发展。其中的一个领域与结构方程模型（structural equation modeling）有关。该模型在建构初期更像是一种对概念加以界定和说明的描述型理论。在界定了概念之后，研究者会进一步对它们之间的关系作出最好的猜测；接着再以图示的形式将概念之间的联系刻画出来。然后，分别测量每一个概念，并对统计量进行有关计算。计算结果揭示了概念之间联系的权重，并能够说明哪一对概念之间的关系最为重要。通过这种方式，调查者可以了解到这些概念在数量上是如何相互联系的。

量化型理论蓬勃发展的第二个领域存在于联结主义（connectionism）[有时也称为并行分布加工（parallel distributed processing）]学说之中。在开始时，此类理论使用类比模型对人类神经系统加以类比说明。在计算机中的单元至少存在三个水平。这与神经元存在三个水平比较类似，并且它们之间还能彼此传递信息。在感知世界的过程中（例如，它们可能会分析字母的曲直），这些单元会发出信号并将其传递到更深层的单元中，使其中一些深层的单元或多或少地发出自己的信号。这一方式与神经元的工作原理是一样的。在经历了一段时间的重复之后，一些单元会开始建立权重，这反映了学习是如何取得进展的（例如，深层单位开始识别一些字母）。这些量化的权重系数可以被理解为一种反映人类神经系统工作的理论表征，而由理论表征所得到的行为结果可以通过量化的方式与人类的行为进行比较。例如，计算机系统是不是跟人们一样容易把一些特定的字母弄错？除了这些例外的情况，心理学中的多数理论仍是描述型与类比型的。然而，我们在努力学习应对由变异及测量尺度所引发的问题，随着心理学逐渐精确化、复杂化，在未来将会有越来越多的量化型心理学理论出现。

良好理论的特性

当你看到一个理论时，你如何知道它是否是一个良好的理论呢？我曾经暗示，量化型理论比类比型理论更好，而类比型理论比描述型理论更好。为什么会这样呢？

首先，一个理论必须能对已经收集到的最多的数据作出解释（account for the most of the data）。提出一个没有数据支持的理论毫无意义。（你可以看出为什么进行一次全面的文献搜索是如此的重要；它能够帮助你在收集数据之前就排除掉一些富有竞争力的理论。）然而，除非有替代性的理论能对证据加以解释，否则通常情况下一两个反证并不足以推翻理论。

一个理论必须具有可验证性（testable）。正如本章开头所述，通过研究可以排除一些理论，但同时又可以支持另一些理论，就在这一过程中，科学得以发展进步。因而，具有可验证性意味着理论可能会被证明是错误的。如果实验结果与理论预期相反，那么，理论就会遭到否定。如果一个理论的普遍性很强，以至于它能对任何一

项实验结果都作出解释，那么，要证明它错误就几乎不可能了①。理论有可能无法被验证，其原因之一在于，这些理论所预测的结果只会在某些时候以一种不可预期的方式出现。譬如，通常意义上所说的 Freud 的压抑理论，它在本质上是无法被验证的。你如何通过实验来反证压抑这种理论呢？或许你可以找一些希望忘记自己某次经历的人。比如说，你可以设法让人们作弊，接着让他们去面对那些屈从于其恶劣行径的人②。有时，一段时间之后，你还可以再让欺骗者的好友私下里去询问他们是否作弊了。如果没有人承认自己作弊，那么，你就为压抑理论提供了佐证，因为这表明，每个人都会在事件发生之后展示出压抑心理(或许也可能是他们撒谎了)。反之，倘若每个人都承认自己作弊了，这一结果并不能推翻理论。这是由于该理论从来都没有声称，在一个特殊事件出现之后，所有人都会出现压抑；只是说有的人有时在某些事件发生之后会表现出压抑行为。如此看来，你的实验丝毫也没有排除该理论。从科学的视角来看，一项普遍到无法通过检验来加以置疑的理论是毫无用途的。

尽管理论不能够对任何行为都作出解释，同时它也不能过于局限。也就是说，理论能解释的可以被直接观察到的事件越少，那么，它就越没有价值。在最极端的情况下，一个理论可能只是对可观测事件之间关系的重新表述而已③。例如，“一个8岁儿童在看了鹏鸟(road runner)动画之后，击打吊袋的次数比以前更多了”的表述不如“观看电视上的暴力节目会导致儿童的攻击行为”的表述更具有实用价值。而比它们更有实用性的表述则是“观看暴力行为会使人们的攻击行为增多。”表述越具有普遍性，那么，它们就越有价值。这是因为它们能对更多的可观测事件做出解释。

一个优秀的理论也应当具有简约性(parsimony)，即在能够解释数据结果的情况下，理论越简单越好。太过详尽和复杂的理论没有什么用处，因为当我们试着将理论应用于新的情境中时，我们往往并不了解所有符合理论条件的情境状态。

一个优秀的理论还应该可以预测(predicts)未来实验的结果。即使是描述型理论也应该能够详细地说明各类事件之间的关系。这样一来，各类可直接观察的事件之间的关系都可以由理论预测。类比型理论和量化型理论也可以对事件间的关系作出预测，并且它们的预期结果更加精确。

最后，最好的理论有助于我们解答一些终极性问题，而不单纯是一些临时性问题。终极性问题(ultimate question)是一个有关为什么(why)的问题。而临时性问题(proximate question)是一个有关怎么样(how)的问题。我们在本章中已提出的理论“儿童观察到的暴力行为越多，那么，他们越有可能出现攻击行为”，它所解答的是一个是什么(what)的问题。然而，它并没有回答最终的为什么(why)的问题：为什么儿童会这样表现？一个基于终极性问题所建立的理论可能会作出这样解释，“演化深深地影响了人类，并促使人们以攻击水平的增加来应对环境中的暴力行为，这是由于该适应性使人类更有可能生存下去。”现在心理学中的绝大多数理论

---

① 出于这一原因，可验证性被一些人称为可证伪性(falsifiability)。正如在本章前面所提到的，Karl Popper (1968) 以提出“构想只有在可以被证伪的时候才具有可验证性”而闻名。

② 让我们暂时先不考虑这个实验是否符合道德规范。

③ 事实上，这样的表述不符合我们对理论的界定，不过一些调查者可能还会将其称为理论。

都是基于临时性问题而建立起来的。不过，最近在心理学领域中，人们对演化理论的重视日益增加，这可能将导致更多能解答终极性问题的理论产生。



一个优秀的理论允许你进行预测……

理论总是比数据先行吗

在下面关于理论与实验之间关系的略带理想化的探讨中，我可能会使你相信，在进行实验之前，你必须先在头脑中构建一个理论。不过，对于某些类型的研究而言，理论可能并不是那么重要。有些研究者更愿意在收集了大量数据之后再建立理论。他们就如同大侦探夏洛克·福尔摩斯（Sherlock Holmes）一样，只有在收集了所有线索（数据）之后，他们才会查出并揭露罪犯（理论）。他们认为，特别是在研究项目开展的初期，在收集数据之前提出理论就像先指定坏人，然后，再收集与其罪行有关的线索一样。这两种程序都有失偏颇。事实上，操作性条件研究的创始人 B. F. Skinner 主张，大多数理论都是弊大于利（Skinner, 1950）。

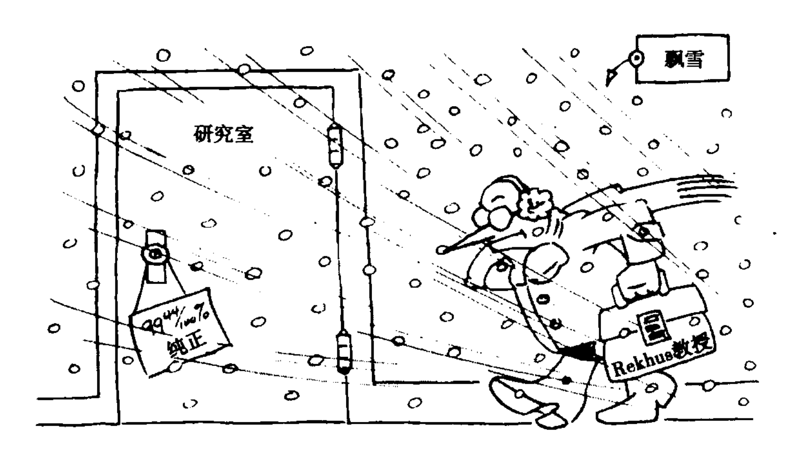
Skinner 相信，我们从事科学研究工作的目的是为了解释可观测的事件。然而，由于理论所使用的是抽象概念而非具体事件，所以，它们往往并不能为我们提供任何帮助。另外，由于理论本身是抽象的，所以，它还会让我们在没有完成研究的情况下，便冒然相信自己的研究已经比较完善了。由此，我们就会在并不清楚理论是否正确的情况下就试图使用理论去填补研究的漏洞。最后，Skinner 还担心，当我们使用理论引导自己的研究，而理论却被证明为错误的时候，我们会失去许多由该理论所得出的研究结果。要理解 Skinner 相当极端的立场，就必须将其放入行为主义学说背景之中，该学说对所有心理事件和作为许多现代理论基石的中介变量都怀有一种普遍的排斥感。如今大部分做研究的心理学家都不认同 Skinner 的观点，他们认为，理论对于绝大多数类型的研究都是至关重要的。然而，即使是这些研究者也明白，有时候理论在引导研究方面的作用并不大。

当一项特定的研究主题尚处于酝酿期时，不成熟的理论所引发问题也许比它能够解决的问题要多。如果在收集大量的数据之前我们就提出了一个站不住脚的理论，那么，接下来我们就不得不花费许多时间和精力去检验这个理论。然而，倘若此时我们将精力集中于收集更多的数据，那么，这可能有助于我们产生一个更好的理论。特别是在一个新研究项目开展之初，单纯地用实验检验这样的问题“我想知道，如果……则会发生什么”比检验一个理论假设更为合适。而当我们进行了充分观察之后，就应当建构理论，因为在这种情况下，以没有结构的探索性方式去收集数据会变得越来越事倍功半。

玩理论游戏的另一个危险在于它的智慧趣味性。一开始，研究者发现了一个重要的问题，进行了初步的观察，并提出一个理论。接着，别人检验了这一理论，推翻了它，并提出了自己的理论；而在其后，又有人推翻了第二个理论，如此不断往复。当理论能引发理论的时候，理论游戏的生命就会延续下去，并且有时候我们会忘记

最重要的问题是什么。最终我们将以调查那些易于检验的理论而结束，而不是去寻找重要的问题。由于没有任何一种科学能够为人们提供考查所有问题所需的任何资源，因此，我们必须要加以甄选。很显然，我们应该去选择重要的问题来研究。如果出于对理论的考虑，我们选择了研究那些不重要的简单问题，而非那些重要的难题，那么，我们就是滥用了自己的资源①。为了防止受到坑害，我多次发誓去讲授编号为心理学 371（Psy 371）的课程，这门课的名称是“心理学所不知晓的事情”。其意图在于，寻找人类行为领域中目前尚无人研究的重要问题，并进一步指出在这些领域中开展研究的策略。

在理论中，最后一种研究类型有时并不那么重要，此类研究被称为应用研究（applied research）。应用研究用来解决特殊的问题，这与基础研究（basic research）恰好相反②，后者唯一的目的在于丰富科学领域的知识。就这一点而言，现在的大部分研究都是基础研究。尽管基础研究并不是用来解决实际问题的，但它有助于我们解决此类问题。能够提供某些最有效的程序、纠正人类行为问题的现代行为矫正技术就有赖于实验室中以老鼠为研究对象所开展的基础研究。Jack Adams（1972）发现，在 20 世纪 70 年代，许多军事系统设计所使用的信息便来源于 20 多年前的基础研究。



另一方面，应用研究的主要目的在于解决问题。或许你需要知道人们是如何阅读手写数字的，由此你便可以设计出阅读邮编的机器。或者你希望了解日常测验能否改善学生在重要考试中的成绩。或者你还想知道，认知行为疗法是否比精神分析更为有效。许多这样的实际问题都需要得到解答。然而，这可能是基础研究永远无法做到的。从事研究的一个非常合理的原因是对实践中出现的问题给出令人满意

---

① Kuhn 在其《科学革命的结构》（1970，37 页）一书中指出，事实上，一旦科学团体接受了一种范式（一组假设，被广泛接受的模型，或综合性的理论），科学家们就会只去从事那些能够在该范式下解决的问题：“在很大程度上，这些才是科学团体所承认的问题……就此而言，一种范式甚至可以使科学团体与那些不能被还原为谜题形式的重要的社会问题完全绝缘，这是因为那些问题无法以范式所提供的概念和工具来表达。”

② 基础研究有被称为纯粹研究（pure research），这或许是因为从事该研究的工作者一般都没有多种复杂的动机。然而令人遗憾的是，有些从事此类研究的人更喜欢选择字典中的另一种解释，即纯洁的（pure）——例如，没有被邪恶和罪恶污染。我从来都没有听说过纯科学家曾为自己辩驳说他们在身体上是贞洁的，尽管我怀疑这是科学家穿着白色实验服的潜意识原因。

的答案，特别是当你通过研究发现，对世界造成直接影响的时候。在许多情况下，应用研究也同样有可能验证一个理论。此时，研究不仅能带来直接的实际贡献，而且还有助于我们建构科学的知识体系。

观察也是产生应用研究构想的关键要素。寻找实际问题只不过就是认真地观察人类行为，并充分满足你的好奇心。同时借助我们先前所讨论的那些有助于获取实验构想的其他程序，你将会发现，需要解答的实际问题要多于你可以通过实验来解答的问题。所以，如前所述，问题不在于“我能做什么？”而在于“我首先应该做什么？”

心理学研究的重要性

在本章结束之前，我还想指出，尽管我强调从事研究的趣味性以及它能满足我们对于人类行为的好奇，但是，开展心理学研究是必须的，因为它可以解答我们生活中一些最重要的问题。如果我让你找出对社会造成最大不幸并且耗费我们最多金钱的问题是什么时，你会说出哪些呢？以下是我能想到的10个问题：

1. 我们的孩子在学校里学的知识技能不够多（例如，许多人都不能阅读、写字、解决数学问题等）。
2. 有太多的人滥用药物。
3. 人们的行为方式不健康（吸烟、传播性病、饮食欠佳、不做运动等）。
4. 人类冲突引发战争。
5. 家庭暴力伤及家庭成员。
6. 暴力行为导致高犯罪率。
7. 有太多的人接受福利救济。
8. 社会文明遭到破坏，比如，驾车司机的暴力行为、乱丢垃圾、粗鲁无礼、轻率的诉讼等。
9. 有太多的人因意外事故而死亡或受伤。
10. 许多工人因训练不足而无法胜任工作，他们仍需要再培训。

这些问题中有多少是有关人类行为的问题，又有多少问题处于心理学研究范畴之中呢？你回答对了——所有问题都是！想一想政客们谈论的问题，如犯罪、健康、教育、毒品、经济。这些领域中的问题因人们的行为方式而产生或改变。为了解决这些问题，我们需要更好地理解人类的行为，而要理解人类的行为就需要开展研究。因此，作为心理研究者，我们所做的工作不仅是出于兴趣和娱乐，而且也源于工作本身的极端重要性。社会需要这样的研究，而研究成果也有助于人们更好地生活。

小　结

尽管我们对人类的行为都有一种天生的好奇心，然而，我们中有许多人都因为一些非理性的恐惧阻碍了新想法的产生。一些人担心其他所有研究者都是天才，认为自己的想法不具有原创性。还有一些人害怕提出需要使用复杂仪器的实验，而另

外一些人则害怕提出只使用简单仪器的实验。有的人害怕自己提出的构想太简单,担心自己的实验需要借助复杂的统计来完成,或害怕自己提出的构想不够完美。此外,还有些人如果不将自己的构想转化为科学术语,就不敢相信自己提出了好的想法。

获得实验构想的关键在于学会观察(observe)周围的世界。你还需要知道,从科学的视角来说,哪个构想更为适宜。要成为实验构想,这些想法就必须是可重复的(repeatable)、可观测的(observable)和可验证的(testable)。你可以通过观察自己、朋友、儿童,甚至宠物来产生研究构想。尽管一些最好的想法来自于直接观察,不过你也可以通过阅读别人的研究(替代观察)以及深入分析自己的研究来获得构想。

将观察转化为实验的典型方式是形成理论(theory)。理论是有关一组抽象变量之间可能存在的关系的陈述。通过观察引出理论的过程被称为归纳法(induction),即由特殊事件得出一般性陈述。实验中的预测被称为假设(hypothesis),它可以通过演绎法(deduction)的方式由理论推出。如果检验这一假设的实验被证实,那么,理论就得到了支持,但不能说得到了证实。而当假设被证明为错误的时候,根据逻辑规则,就可以说理论经过演绎推理被证伪了。然而,由于推翻假设所使用的统计测验具有偶然性,因此,否定理论的证据并不都是可靠的。

描述型理论(descriptive theory)就是给事件命名,与命名具体事件相比,它在命名那些抽象变量时显得最为有用。类比型理论(analogical theories)通过心理联系与物理模型之间的类比说明关系的运作方式。量化型理论(quantitative theories)则以数学术语详细地说明事物之间的关系。心理学中很少用到量化型理论,这是因为我们仍在学习如何解释变异性以及如何开发精确的测量工具。优秀的理论能够解释大多数数据(accounts for most of the data)、具有可验证性(testable)、限制性弱(not too restrictive)、遵循简约原则(parsimony),能够预测(able to predict)未来实验结果,而最好的理论有助于解答终极性问题(ultimate questions)[为什么(why)的问题],而不仅仅是临时性问题(proximate questions)[是什么(what)的问题]。

Skinner 认为,实际上理论是没有用的,因为它们无助于解释可观测的事件。同时理论还会误导我们相信自己已经做完了那些尚未完成的实验,而且当理论被推翻时,研究也就变得毫无价值。尽管大多数研究者不认同这一观点,他们相信理论是有用的。然而,理论也不是所有研究所必须的,特别是在研究的初级阶段。应用研究(applied research)便是如此,它是用来解决问题的。这恰好与基础研究(basic research)相反,因为基础研究可以丰富科学知识体系。

4

如何公正地对待参与者

像你希望别人如何待你那样对待别人。

MATTHEW 7:12

我们的数据结果表明,竞争与奖励的社会结构是以人为被试的实验中随意行为的来源之一;为了获得别人认可,平庸的科学家最有可能随意而行。

B. BARBER(1976)

噢,当我们第一次尝试撒谎时,我们编织了一张错综复杂的网。

SIR WALTER SCOTT

除了对活体动物进行实验和观察外,人们没有其他方法去探寻有机界的规律。

IVAN P. PAVLOV

不管我们是两只脚站立,还是四只脚,或者根本没有脚,从伦理的角度看,我们所处的地位是相同的。

P. SINGER(1985)

既然你已经有了一个实验想法,那你就准备开始仔细筹划实验。首先,不管用哪种方法,我们都要考虑伦理问题。作为实验者,我们至少会在两个方面有悖于伦理。我们可能虐待我们正在测量其行为的人或者动物;我们同样可能虐待我们尝试建立的知识体系——或者说是,不公正地对待科学。在这一章,我们将讨论公正地对待参与者;下一章讨论公正地对待科学。

社会作为一个整体,它拥有一系列规则;科学界尤其如此,我们必须遵照这些规则作研究。一些规则是约定俗成的,如基本礼貌规则等。一个假定是这些规则是显而易见的,每个人都应该明白。其他规则是以书面形式表达出来的,如《心理学家的伦理原则和行为准则》[American Psychological Association(APA),2002]以及《研究人类参与者的伦理》(Sales & Folkman,2000)。这些规则随着实验的社会角色概念与个体权力的变化而不断被修订。在本章的第一部分中,我们将讨论实验者和参与者之间的关系,其中包括基本礼貌问题。然后,我们探讨这一关系如何影响实验结果以及实验者—参与者两种不同类型的关系。最后,我们再讨论公正对待动物的伦理问题。

公正对待人类参与者

心理学研究的目的是理解行为,所以,我们经常与人打交道(有些情况下是动物)。传统上,心理学家把提供行为的人称为被试(subjects)。早期的心理学开拓者喜欢这个词,因为它听上去很科学,而且那时的研究被试是人类。不幸的是,这个词也暗含受试者可能会受实验者意愿的影响,或者更为糟糕的是受其支配。早在20世纪30年代就有人建议用被实验者(experimentee)一词代替被试(subject),但是这一建议并没有得到认可(Rosenzweig, 1970)。

那么,应该用什么词?这种讨论看上去有些琐碎。在某种情况下,被试(subject)这个词反映了个人和实验者之间关系的本质,并且暗含了特定的伦理意向。被试是被动的,对实验的条件进行反应,就像实验室中的化学物质被放在一起产生反应一样。基于这些原因,在1994年美国心理协会建议改变这一术语,将被试(subject)改为参与者(participant)。美国心理学会认为,这个词恰当地反映了参与者通过参加研究而给予我们的帮助,并且为参与者界定了与实验者平等的地位。正如你会在13章看到的,在写研究报告时我们最好称呼参与者的身份为学生、儿童、妇女等,但最合适的通用词还是参与者。

用参与者(participants)而不是被试(subjects)并没有得到普遍接受。例如,心理学研究学会(Psychonomic Society)允许作者不参照这一规则。Roddy Roediger,美国心理学会[American Psychological Society,现在是心理科学协会(Association for Psychological Sciences)]的前任主席,强烈反对用参与者(participant)代替被试(subject)一词,并声称他给美国心理协会杂志投的文章有特许权,因为他在给美国心理协会的一个文字编辑的一封信中描述了一个微妙的情况,现摘录如下:

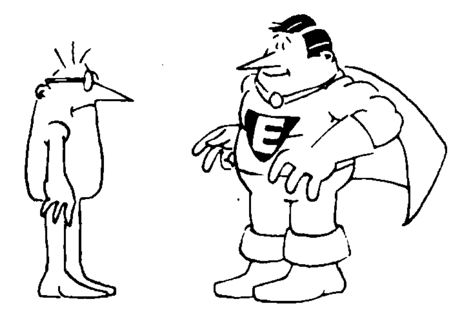
> 我必须指出,我是参与恐惧症(Sufferers of Participant Phobia)(SPP)患者中的一员(事实上是发现者)。因为我必须在APA许多杂志上使用参与者这个词……我同样是参与恐惧症幸存者(Participant Phobia Syndrome

Survivors)(PPSS)中一员。在我们的杂志中使用参与者这个词会导致我精神痛苦，产生过度的压力，这促使我写了这封信，并向一个支持小组寻求帮助(那些认为语言会改变厌恶的实验心理学家……)。(Roediger,2004)

很明显，这封信用一点带有挖苦的幽默夸大了事实，它确实表达了一些研究者的担心；虽然这些担心不是不无道理，但它只是一些压力群体出于某种政治目的而不是科学研究而产生的过度反应。

在实验心理学的早期历史中，没有人担心如何称呼做实验的人，因为实验者和参与者是同一个人。在那个时候，大部分心理学家将自己所报告的内部经验作为实验的因变量。心理学家认为，只有通过时间和训练才能意识到那些内部经验，所以，他们认为自己是最好的参与者。

后来，许多实验者开始意识到，通过内部事件的口语报告所获得的数据不符合心理学的科学性。他们认为，客观性和主观性是不可能同时存在的，这些心理学家开始了心理学的一场革命。一些对心理变革反应过强的心理学家认为，只有动物才适合心理学实验。如果有人意识到参与者的口语报告不合适的问题，那么，就应该选择不会说话的动物①！在这一时期，小白鼠成了实验的主要参与者。其他研究者认为，尽管实验者做实验的经验很丰富，但小白鼠与人类大不相同。实验所需要的是一个真正的人，他们选择的是大学生。在已发表的研究中有70%到85%都是大学生参与者(Schultz ,1969; Smart, 1966)，并且有90%的研究是由大学的心理学系完成的(Jung,1969)。



98磅重的参与者　　超级实验者

根据最新的观点，参与者是指能对实验操作做出真实反映的，诚实的、动机较强的人。然而，参与者并不是想法单纯的观察者。他们通常对自己参加的实验有明确的想法，并且尝试获得特定的目标，这些目标通常与实验者的目标不同。

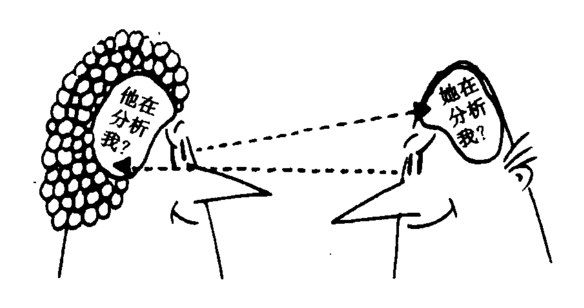
人类(甚至是大学生)同样有特定的法律和道德权利。物理学家可以从倾斜平面实验中拿出木块，扔掉它，敲打它，骂它，亲吻它，或者做一些其他的事情。尽管他的同事可能认为他相当怪异，但并不能逮捕他或将他从他的职业中赶出来。但是，心理学家必须时时刻刻尊重参与者的权利。

实验者—参与者关系的本质使得参与者相当的脆弱，因为通常实验者拥有大部分权利。例如，许多人做实验是因为心理课程的要求。在这种情况下，学生会认为，如果他们不能完成实验者所要求的，他们的课程成绩就会受到影响。另一方面，如果给被试费，那么，他们会以为，不与实验者合作，他们得到的钱就会减少。此外，如果是自愿来做实验，由于他们认为，自己能够促进心理科学的发展，所以，他们觉得社会可以从其合作行为中获益。无论哪种情况，参与者都会觉得实验者有最终的评价或操纵他们行为的权利。

参与者与实验者的合作除了以上的学术、金钱或利他动机外，他们通常还具有

① 好的，我对历史采取了一些自由的态度。这章后面我将给出在试验中使用动物的更好的理由。

一种普遍的观点，即实验者有一种神秘的窍门能够知道他们是否合作。心理学家和陌生人交谈的头三句话就能阐释这种观点："你的职业是什么？""我是一个心理学家。""哦，你能分析一下我吗？"出于种种原因，许多人认为，每个心理学家都有一个类似于X射线的视力，他们能看穿人们的心灵，知道人们在想什么。人们觉得自己最好合作，否则实验者就会知道。这一种神秘感再次使得实验者—参与者的关系有利于实验者。



礼貌规则

为了缓解这种关系，实验心理学家需要遵守一些行为准则，使参与者获得尊重和尊严。

作为一个实验新手，你应该在你的实验室挂一副标记（假想的也行），上面写着："参与者也是人！"像对待在课题上给你提供帮助的人那样去礼貌地对待参与者。你应该遵守一些简单的礼貌规则。

1. **确保出席**。实验者经常会忘记有参与者已经报名参加实验了，或者没有考虑到实验仪器受损、实验推迟及取消后参与者的感受。一旦有人报名参加实验，你应该尽最大的努力完成你的责任，出席实验。
2. **准时**。参与者的时间也非常宝贵，不要浪费它。
3. **准备**。在你见到所有参与者之前应该熟练了解实验的每个阶段。做其他事情不仅不礼貌，而且如果你在给指导语时结结巴巴，还要考虑如何使用设备，在实验时笨手笨脚、嘟嘟囔囔，那么，参与者就会很疑惑或者厌恶，他们的表现就会很差。
4. **礼貌**。除非实验任务要求，否则，你应该请求参与者做某事，而不是命令他们。经常使用"请""谢谢"和"不客气"。
5. **保密**。对实验中参与者的信息保密。不仅对参与者的言语保密，而且也要对参与者在实验中的表现保密。联邦政府的捐赠与你获得的信息、你如何使用这些信息以及你如何编码和存储这些信息有关。如果条件允许，把参与者的名字从数据表上删去，并使用一些方法使其他人无法知道参与者的身份。
6. **专业性**。你不要过分的严肃和呆板，这会使参与者感觉不舒服；也不要太随意或轻率，以至于参与者会觉得你不是很重视实验，这样参与者也不会重视实验！实验不适合用来约会、寻找高尔夫伙伴、卖保险，实验者—参与者关系也不能用来达到研究以外的其他任何目的。

这些规则看上去非常简单，但并不是所有的伦理问题都这么直接关注人类参与者。一些复杂的问题，如"什么是知情权？"和"是否允许有心理压力？"在《研究人类

参与者行为的道德原则》(*Ethical Principles in the Conduct of Research with Human Participants*)(APA,2002)中有详细阐述。但是,没有出版物可以包括所有的伦理问题,当要求实验者提供不偏不倚的意见时,他们会采用“结束通话”的方法。由于这些原因,面向研究的机构是机构审查委员会(institutional review boards),简称IRBs①。这些IRBs是由经验丰富的研究者,有时是医生和其他技术专家组成的。所有使用人类参与者的研究都要受到这些委员会的审查②。通常,研究者要填写一个表格,表格包含一些问题,如:

“参与者是否要求有知情权?”和“能否保证数据的保密性?”这类表格还要求研究者对研究进行简述。委员会的成员特别重视可能给参与者的生理或心理带来伤害的问题。在任何研究中期望伤害减少到零都是不现实的,参与者甚至可能在地毯上绊倒而摔坏一条腿。但是,审查委员会的任务是确保伤害的风险最小化。所以,既然知道风险是研究中的一部分,委员会的任务就是确定从研究中获得的益处是否大于这些风险。

审查委员会必然会帮助消除或改善许多可能不符合伦理道德的研究。但是,在生物医学领域,IRBs自己却成了研究对象。因为有人发现委员会中有相当多的人不擅长平衡人类研究的风险和益处(B. Barber,1976)。接受调查的委员会成员大部分没有受到过任何正式的、伦理研究的培训。

一些心理学家认为,很少有证据证明IRBs能够有效地减少对人类参与者的风险(Mueller & Furedy,2001)。还有人认为,IRBs过于吹毛求疵,以至于他们侵犯了社会科学中的知识产权调查。负责监督IRBs的联邦政府办公室的前任主任Tom Puglisi(2001)等人认为,IRBs服务于一个必要的目的,并且适当的参照法规能使更多的社会和行为科学研究避免法律纠纷。换句话说,当大部分心理学家将他们的研究计划上交给IRBs时,他们应该陈述自己的研究如何能不与法规相悖,而不是试图证明自己的研究。无论你对IRBs的作用如何看待,你应该知道,研究的伦理责任,最终责任应该由实验者承担。

知情权

IRBs和你都关心的问题之一是知情权(informed consent)。在参与者答应参加实验之前,参与者有权知道那些可能影响他们作决定的一些因素。一旦告知参与者之后,研究者就应该以书面形式将它们记录下来。虽然告知参与者及获得他们的同意看上去非常直接,但是许多因素会掩盖这一问题。如文件上记录的信息易于理解,保证处于从属地位的参与者不是被迫参加实验,或者有些参与者是否有能力作出正式的决定。《心理学家的伦理原则和行为准则》(American Psychological Associ-

① 国家健康机构要求它们资助的所有研究都有IRBs。它们出版了一项方针“保护人类研究主题:机构审查委员会指导”(NIH Guide, Volume 22, Number 29, August 13, 1993)来帮助审查委员会成员、研究者和机构管理员完成他们的责任。可以从美国政府印刷局获得副本,文件管理员,邮政信箱371954,宾夕法尼亚州匹兹堡15250(参考邮政总局库存号017-040-00525-3)。

② 在一些情况下,如上课所需的书正是你所拥有的一样,老师可以说服IRBs允许他们为考虑伦理因素来评价课堂实验。当说服了IRBs,老师们就会考虑哪些行为是允许的。虽然,有时课堂实验是很重要的,足以发表,但大部分实验的首要目的是为了教学生做实验。大部分情况下,教学生做实验可以通过低风险实验或高风险实验来完成。所以,如果当你准备一个课堂实验,参与者带来了一些毒品并且向观众展示他们的性幻想……不要理会他们!

ation,2002)对这一问题进行了比较详细的说明:

8.02 研究的知情权

(1)如果心理学家希望以标准3.10获得知情权,那么,他们应该告知参与者如下的知情权:

(a)研究的目的、预计期限以及过程;

(b)参与者有权参加,并且即使在研究开始后也有权退出研究;

(c)参与或退出研究的可预见的后果;

(d)影响他们是否愿意参加实验的一些可预见的合理因素,如潜在风险、不舒适感或者负面影响;

(e)任何预期的研究益处;

(f)保密的限制;

(g)参加实验的报酬;

(h)解决研究和研究中参与者权利问题的联系人,他们能够解答参与者的问题。

“伦理原则”同样阐述了在哪些情况下不需要知情权,但是,你仍应该得到IRBs的同意,以确保你的解释是准确的①。

8.05 研究不需要知情权

只有在以下两种情况下,心理学家才可以免除知情权。

(1)研究确保不会带来困扰或者伤害的情况下,包括:

(a)在教育环境中进行的正常教育实践,教学内容或者课堂管理方法的研究;

(b)只有那些公开参与者的反应不会给参与者带来刑事或民事法律责任,不会破坏他们的财务状况、就业、名声以及保护其秘密的匿名问卷,自然观察,或者档案研究法;

(c)研究那些在组织背景下进行研究、工作或管理效率有关的因素,研究不会影响参与者的就业,保护其秘密,或者

(2)在法律或者联邦、机构法规允许的情况下。

在你知道了研究的方法、参与者了解知情权之后,你应该考虑你的实验中实验者—参与者关系的本质。这一关系的本质非常重要,因为它不仅影响参与者的权利,也影响实验的结果。尽管实验心理学家倾向于认为实验心理学中的参与者是那些对无菌、受控制的环境进行反应的自然生物,大多数人都知道情况并非如此。下一部分我们将更加详细地探讨实验环境是如何影响实验结果的。

需求特征

当参与者来做实验时,他们不知道自己会被要求做什么。但是,他们通常对实验感兴趣,并希望明确了解实验目的。反过来,实验者通常对自己的实验目的进行

---

① 当1b是正确的,即没有预期的伤害并使用匿名问卷,那么,就不需要知情权,IRBs对这一条有些质疑。在委员会会议上我看到人们对这一条的争论比其他条多。

保密,这更加推动参与者通过实验者给他们的线索去探究实验的目的。于是,实验就成了一个解决问题的游戏。

在实验环境中影响参与者的这些线索被称为需求特征(demand characteristics),因为它们要求特定的反应(Orne,1962)。尽管实验者给参与者很多线索,参与者仍然有需求特征。如果参与者学过心理课程,阅读过心理实验,或者仅仅由朋友告诉他们有关实验的内容,他们就会产生如下期待:实验者会使我震惊,实验者试图发现我有多聪明,实验者诱使我透露有关自己不光彩的事情。

有时这些想法十分强烈,以至于参与者不能从中摆脱出来。在我的一个实验中,参与者被要求记住一组通过耳机呈现的词。刚开始实验不久,他扔掉耳机,喊道"这东西吓到了我!"我想他可能是对的,所以,我仔细地检查从耳机中传出的电流。这个耳机是好的。我尝试继续实验,但是这个人仍然声称他受到了惊吓。他非常肯定地认为,我准备伤害他,并且不相信任何解释。最后,他的数据必须去掉。

其他的需求特征来源于实验中参与者获得的微妙线索。为了减少这些线索,实验者试图对实验过程进行标准化。例如,实验者通常宣读写在一张纸上的指导语,以保证所有参与者至少有相同的口头表达要求特点。但是,在一些实验中,实验者宣读指导语的方式甚至也会影响参与者的反应。在一个实验中,实验者做了两组录音的指导语,他希望产生相反的实验结果(Adair & Epstein,1968)。实验者发现,听不同录音指导语的参与者的表现有显著性差异。尽管实验者读的是相同的指导语,但不同实验者声音之间的细微差异使得结果明显地符合他们的预期。

甚至动物也会受到实验者给出的微妙线索的影响。有这样一个著名的实验者偏差实验。实验者是学生,他们训练小白鼠走迷宫(Rosenthal & Fode,1973)。一些实验者被告知他们的小白鼠经过特殊地培育变得非常聪明,学习速度快;另一些实验者被告知他们的小白鼠经过培训变得愚钝,学习速度慢。尽管那些被认为是聪明的小白鼠事实上与那些被认为是愚钝的小白鼠是一窝的,但那些被认为是聪明的小白鼠用更少的次数就学会了走迷宫。对这一结果的通常解释是学生对待小白鼠的方式不同,他们与"聪明"的小白鼠玩的次数更多,以至于小白鼠在实验中受操作的恐惧感减少。但是,其他研究者认为,这一结果是学生捏造了数据造成(Barber & Silver,1968)。无论什么原因,实验的结果反映了实验者的偏见。

尽管我对需求特征的描述使得这一概念听上去相当不好,但它可能没有我想象的那么糟糕。研究者(Weber & Cook,1972)发现,很少有证据表明,参与者通常会尝试去证实他们相信的东西就是实验者的假设,而实验者的这些假设是参与者从实验中推断出来的。相反,有些研究者认为,参与者在表现自己最好的一面,即他们试着表现出能干、正常和可爱。参与者认为,别人如何评价他们比他们如何完成参与者的期待或者证实那些假设更重要。

T. X. Barber(1976)在一本关于解决人类研究缺陷的书中描述,许多研究者认为,需求特征本身受到其他设计缺陷的严重削弱。他认为,支持这一观点的许多实验都是糟糕的。但是,正是因为这些研究可能存在缺陷,因此,我们不能推断需求特征可以作为实验中的潜在问题而忽视它。我们可以使用任何能够减少需求特征潜在效应的方法,以便改善实验。

参与者对需求特征的反应

如果参与者在一个实验中觉察到了需求特征,他们可能作出什么样的反应?

合作型参与者。在人类参与者觉察到实验的需求特征后,他们的反应就可能取决于他们对待实验的态度(Adair,1973)。大部分人都会合作并试着去完成参与者的要求。有些合作行为达到了惊人的程度。在一个调查合作性的实验中,实验者给1个参与者一叠纸,有2 000张,并让他完成每张纸中的224道加法问题。尽管这一任务很明显是不可能的,但这个人一直做了6个半小时,这样的合作程度令实验者几乎崩溃。第二个实验中,实验者要求参与者在完成加法运算后,再把每张纸都至少撕成32片。同样,他们在这项任务上也一直坚持好几个小时,而且也没有表现出反对的态度。

为了了解参与者对需求特征反应背后的合作行为,请看以下有关群体压力的实验。实验将1个人和其他6个人一同带进一个房间,并给这一群人呈现一些问题,要求他们判断两条线段中哪一条更长。刚开始的几个问题都很容易,大家的看法都一致。然后,再呈现两条线段,参与者肯定上面一条线段更长,但其他人都说下面一条线段更长。经过长时间的踌躇,参与者最后同意下面一条线段更长。这一实验中究竟发生了什么呢?实验者设计这一实验是为了证明群体压力是否会使人作出明显的错误反应。房间里的其他人都是被实验者训练好的同伙或傀儡,并在实验的适当时候撒谎。因为这个真参与者受到了群体压力的影响,所以,实验者认为这一假设得到了证实。但让我们仔细分析一下参与者的想法①,看看究竟发生了什么:"哦,这是另一对线段。明显上面一条长。这是一个多么愚蠢的试验!为什么浪费我们的时间让我们做这么容易的试验呢?为什么我们要一组一起做?实验者一定是想看我们是否会相互影响。当然,每个人都说下面一条线段更长。他们肯定不会真的这么认为。让我想一想,我要么作出让步,同意他们,要么坚持自己的观点。我想很好地完成这一任务,这样我就可以离开这里了。我肯定一群人会使其他人改变自己的观点,我也同意这项实验。另外,实验者看上去像个好人,我不想破坏这一实验。"

如果我们对参与者的心理分析是正确的,那么,实验者的结论就是错误的。参与者只是尝试着合作,但他的合作使我们得出了错误的结论!事实上,直到20世纪70年代,大家才知道在类似于上述实验的顺从研究中,参与者通常都非常多疑,有50%~90%的参与者表现出这种疑虑(Glinski, Glinski,& Slatin,1970)。但是,多疑对行为的影响看上去微不足道。换句话说,多疑的参与者和一般参与者的行为没有多大差别(Kimmel,1996),就算有猜疑效应的时候,参与者也倾向于将自己表现得更好,而不是对实验者采取消极的反应。

防守型参与者。有些参与者对他们自己表现好坏的关注多于对实验者表现好坏的关注,我们称他们为防守型参与者。这些人探索需求特征的方式和合作型参与者一样,但他们使用需求特征的方式不一样。这些参与者经常尽可能出色地完成任

---

① 看,心理学家的确有神秘的力量。

务，这对实验很有价值。但是，在有些实验中，尤其是态度实验，这类参与者就会带来一些问题。

假设我们研究西班牙人和西班牙裔美国白人对待儿童性别角色行为的差异。我们张贴一张登记表，要求自愿者有西班牙姓氏并且母语是西班牙语，第二张登记表要求西班牙裔美国白人，但都不符合第一张登记表中的两个条件的自愿者。现在我们让每个志愿者观看传统性别角色儿童（如女孩子玩玩偶）和非传统性别角色（如男孩子玩玩偶）儿童的图片。然后，我们要求参与者对每个行为的可接受性进行打分。假设更多的西班牙人认为，非传统行为是可以接受的。我们可能推断西班牙人比西班牙裔美国白人更开放。在另一方面，也可能存在另一种解释。每组成员都知道他们是基于种族选择出来的。假设西班牙人比西班牙裔美国白人更加关注民族自豪感，在这种情况下，他们可能委屈自己，生怕自己被认为是社会不能接受的沙文主义者。换句话说，他们可能感受到了实验的需求特征并尝试去保卫自己的民族。

下面是一个真实的实验，它表明了防守型参与者对需求特征的反应。该实验要求参与者用右手食指按键，然后，用左手食指按键（Rosenberg，1969）。利手手指的按键速度通常要快，但第一组参与者被告知耶鲁大学和密歇根大学的研究生两种手指的按键速度相同。第二组参与者没有被告知这一信息。第一组参与者两手指按键的速度差比第二组参与者要小得多。同样，参与者觉察到了实验中不太微妙的需求特征，并试着使自己看起来尽量优秀。

不合作型参与者。还有些参与者既不属于合作型，也不属于防守型，而是彻头彻尾的不合作型！这种行为的结果被生动地称为“修理你效应（screw-you-effect）”（Masling，1966）。不合作的人尝试去发现实验的需求特征。然后，采用能造成与实验者的假设相矛盾的结果的方式行事。这些人的动机可能各不相同。他们参加实验可能是为了完成课程要求并对这种强制性要求不满；或者他们可能反对研究人类行为科学的整体思想；或者可能仅仅是他们不喜欢那个实验者。无论什么原因，这种人对一个实验来说都是相当麻烦的。消除不合作参与者的一种方法是设置行为表现的标准，以便把低于这一标准的参与者的数据剔除。然而，你应该在实验前就确定好这一标准，并在报告实验时说明这一标准。

但是，甚至连这种方法也有可能无法剔除所有不合作参与者的数据。有时，我们所能做的就是让参与者对我们的实验形成一个积极的印象，并且希望他们能够合作。

怎么减少需求特征

尽管我们不能从一个实验中完全消除需求特征，但是，仍应该尽最大可能去减少需求特征。因为它们可能通过影响自变量的水平而成为混淆变量。行为的改变是取决于实验者操纵自变量，还是取决于参与者感受到的需求特征？这个问题非常重要。可以通过以下几种方法来减小需求特征引起的混淆。

自动化。通过实验的自动化尽可能地控制需求特征。我们之前讨论了使用录制好的指导语是一种自动化。不管用何种方法，实验者总是不擅长读指导语，特别

是在大声朗读其 20 次或 120 次以后。如果你想减小由声音影响带来的实验者偏差，你可以找一个不知道实验预期结果的人录制指导语。

我同样也在我自己的实验中使用录制好的指导语，或者使用电脑呈现，包括音频和视频的。在这种情况下，如果试验涉及事件的复杂结果，那么，演示试验可以尽可能慢地呈现以使参与者能够理解，这样就不需要实验者重新解释之前的指导语了。

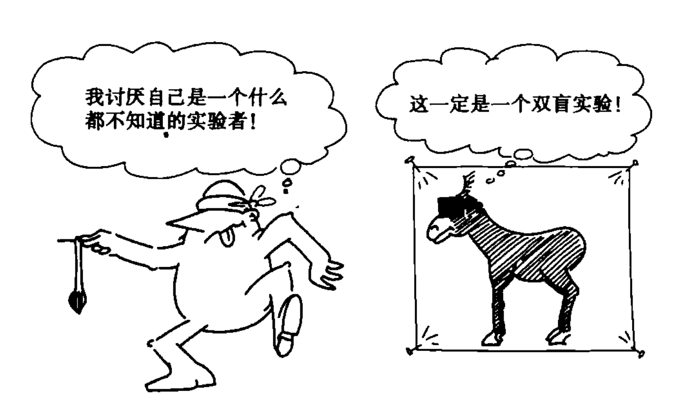
在有些实验室中，使用计算机替代完成全部或部分实验者的作用。有些研究使用电脑上的程序，参与者甚至都没有见过一个实验者。参与者在约定的时间来参加实验。有指示标志告诉参与者坐在电脑终端并按键。然后，电脑演示指导语。参与者表示他或她理解指导语后，便开始实验。使用这一方法背后的基本思想在于，如果参与者并不是我们预想的被动机器人，那么，我们可以使实验者变成机器人。但是，也有研究者反对这一过程，其理由是，这种人工化的环境不仅使参与者觉得缺少人性化，而且影响了研究结果的推广性。这一过程同样要求参与者能够读懂指导语，因此，有些参与者就不适合，如儿童和小白鼠（和大学二年级学生）。

盲和双盲。第二种减少需求特征的方法是不让实验者知道呈现的是自变量哪一个水平。通常情况下，参与者并不知道呈现给他们的自变量水平。因此，这类实验通常称为盲实验。但是，有时既不让参与者知道也不让实验者知道实验的操纵很重要。例如，我曾经做过一个能否用手指来“感受”颜色的实验。实验中，蒙住参与者的眼睛，给他们三张卡片，2 张红的和 1 张蓝的。每一次试验都要求参与者将相同的 2 张卡片放在一堆，不同的那张放在另一堆。我担心当参与者将卡片准确地放置时我会无意识地暗示他们，如通过改变我的呼吸速度，咳嗽或自言自语。我的一些信奉 ESP（extrasensory perception，超感知觉——译者）的朋友甚至说，当参与者反应正确时我可能给他们发送了 ESP 信息！为了避免这些暗示，我站在一个屏幕后面，以确保我看不见参与者。因此，我也不能看见参与者所感受到的颜色。事实上，这一过程有时称为双盲，因为参与者和实验者都不知道呈现给参与者的是哪个自变量水平①。心理药理学家研究药物对行为的影响时，通常使用双盲设计。假设你是一个研究者，希望知道一种刚生产出来，被称为 Crowzac 的药能否治愈那些一看到乌鸦就抑郁的人。你知道，如果你只把这种药给一组病人，然后，观察他们的抑郁是否减轻。这种做法确实存在问题。抑郁症改善可能仅仅是因为参与者的期待，他们认为，这个药能帮助他们。如果你来评价病人的抑郁，你也可能看到一个改善的假象，因为你期待这一改善。为了防范参与者或实验者期待带来的影响，你可以选择另一组人（控制组），但不给他们药物。你必须用对待药物组一样的方法对待控制组，并不真的给控制组这种药，而是用安慰剂来取代药物。

使用安慰剂是指用施加活性药物的方法向参与者施加非活性物质。如果这个药是药丸，那么，安慰剂就可能是糖丸；或者如果需要注射，那么，安慰剂就可能是生理盐水。甚至在研究大麻时，就要设计一种味道很像大麻但是不包含活性成分的安慰剂烟。安慰剂的目的就是产生一个双盲设计，实验者和参与者都不知道哪些人接受的是药物或安慰剂。

---

① 我的一个审稿人认为，这一过程致使我产生了双色盲。



有时不想让参与者和实验者知道呈现给参与者的自变量水平是困难的，甚至是不可能的。如果你是一个研究者，对设备照明情况是如何影响工人生产效率的题感兴趣，你可能让一组工人在现有的照明情况下工作，让另一组工人在照明增强的情况下工作。显然，当工人进入房间的时候，他们就会明显感觉到照明情况，而且你没有办法避免这一点。我使用这个例子是因为这是20世纪20年代做的第一次实验，通过这次实验得出了霍桑效应。霍桑这一词是实验进行的西部电力公司厂房的厂名。研究结果表明，无论照明情况如何，两组工人生产率的增加量相同。于是，霍桑效应是指行为的改变仅仅是取决于实验者对参与者的关注，而不是自变量的影响。这一研究的总体结果是，超过5年的时间，无论工作情况（如照明、休息时间和工作时间）如何改变，工人的生产率不断提高（Roethlisberger，1977）。

最初，研究者对这一结果的解释是，工人的士气随着不同情景中被持续注意而提高。但Mac Parsons（1974）通过回顾了之前的研究发现，在实验过程中工人会得到反馈，即告知他们每天的生产率。这些反馈与工人被注意方式的改变相结合，于是导致生产率的提高。他认为，工人产量的提高是增加强化的结果。所以，霍桑效应的经典解释与霍桑厂房内发生的真实结果并不一致。但是，实验操纵导致行为改变的可能性仍然存在，而不管这些操控是什么。因此，作为实验者，你必须试着去减少这些影响，同时减少由参与者知道他们所接受的自变量水平而带来的影响。即使不能完全消除这些影响，但至少应该清楚它们可能带来的混淆。

多重实验者。解决由实验者引起的需求特征的第三种方法是使用多个实验者。在这种情况下，你就不需要控制实验者变量，只需将很多个实验者随机分配到自变量的各个水平。这一过程可以提高结果的推广性，减少那些单个的、具有明显偏向的实验者对实验结果产生影响的可能性。

需求特征是实验中的一个难题吗

尽管你试图减小需求特征，但它仍然会悄然出现在你的实验。以下是一些用来觉察需求特征的程序。

实验后询问。在反对主观的口语报告之后的很多年中，实验者很少询问参与者在实验之后的印象。幸运的是，现在许多实验者收集了这些信息。这些信息不仅对发现需求特征有价值，而且对提出新的假设也有价值，这些新的假设可以在之后的正式实验中进行检验。

实验后的询问方式有多种,包括从实验者事先无准备的询问到结构化的书面问卷。如果你想确切地发现需求特征,你应该事先准备好问题。

准备问题时应确保这些问题本身没有需求特征。例如,在之前讨论的群体压力实验中,一个具有偏向的问题是"你没有发现其他参与者不是真的参与者,是吗?"这一问题本身要求它回答是。如果答案为"不是",那么,参与者承认他们不是诚实合作的人。而且这种回答也向实验者暗示,这一实验浪费时间,他们的数据没法用。

你要事先设计好问题,使得这些问题从一般的、开放的问题到具体的、具有针对性的问题。例如,设计一个实验研究人们是否会在无意识的情况下形成条件反射,要求参与者谈论一个他们感兴趣的话题,一直说下去直到要求他们停止(Krasner,1958)。每当参与者说一个复数名词时,实验者点头说"很好"或者"嗯嗯",一直给予强化。随着参与者谈话的不断继续,他们会更频繁地使用复数名词。为了证明参与者没有意识到这种条件反射,实验者在实验后问了一个问题:"当你在说的时候你注意到实验者做了什么特别的事吗?"大部分人报告没有。其他一些研究者不相信这一实验结果,做了一个相似的实验,但在询问了那个一般问题后,还问了些更加具体的问题。如"你有没有注意到当你说特定词时,实验者会作出反应?"尽管参与者不能说出是什么词,但大部分人都意识到"当我谈论特定的事情时,如汽车零件,实验者会更加高兴。"那些提到这种意识的人同样也表现出受到条件反射的影响。因为要判断参与者是否受到需求特征的影响,我们应该问他们与需求特征有关的、更为一般的问题。

非实验。判断需求特征是否影响实验结果的另一种方法是设置一个与实验组相匹配的非实验控制组(nonexperiment control group)(Adair,1973)。非实验控制组无需受到自变量的控制。简单地向组员说明实验,提供指导语,展示仪器设备;然后,要求他们估计,如果他们在这样的环境中,将会有何种表现。如果他们的描述与实验组的结果相似,那么,他们可能觉察到了需求特征。导致实验结果的正是这些需求特征,而不是自变量。如果他们的描述与实验组的结果不一致,那么,很可能行为的结果不是需求特征引起的。

例如,Mitchell 和 Richman(1980)对证明心理表象的"准图形"记忆表征的结果提出质疑。在一个经典实验中,实验者要求参与者记忆一个视觉刺激,并生成这一视觉刺激的心理表象。然后,从表象的一点"扫描"到另一点。结果表明,扫描时间与刺激的物理距离存在直接的线性关系。Mitchell 和 Richman 认为,在这一过程中产生了需求特征,因此,他们采用了一个非实验组,仅要求参与者预计他们的扫描时间。这些参与者的结果是一个散点图,与之前的实验结果并不一致。研究者不能消除这种可能性,之前的实验结果可能是需求特征引起的。

模拟控制组。尽管通过询问那些没有参加真实实验的参与者的反应可能对了解需求特征有一定帮助,但这并不能反映出他们的真实行为,可能会误导你。例如,在很长一段时间内,人们对能否要求被催眠的人做出反社会行为或自虐行为感到好奇。1939 年,Rowland 曾报道过一项实验,在实验中告诉被催眠的人,那个很大的莱蛾响尾蛇(响尾蛇的一种)是一根绳子,并要求他们把它捡起来。两个参与者中的一位立刻按照要求去做。实验者用一个隐形的玻璃将他的手挡着,防止他接触到这

条有毒的蛇。但是,在要求42个未被催眠的控制组参与者做相同事情时,有41个参与者拒绝。1952年Young也发现了类似结果(Kihlstrom引用,1995),8个被催眠的参与者有7个尝试去抓放在一个隐形玻璃后面的蛇。另外,他们还愿意将一瓶硝酸泼到实验助理的身上,实验助理同样也在玻璃后面。这些结果能否表明被催眠的人愿意做出反社会和有害的行为呢?

1965年,Orne和Evans设计了一个新的过程,采用模拟控制组来研究这一假设。模拟控制组也接受实验条件,但没有对自变量进行严格的控制。在这种情况下,对那些极容易受催眠的实验组被试进行催眠,然后要求他们抓一种澳大利亚两步蛇。因为这种蛇咬了你之后你只能走两步!所有的参与者都答应这么做,并且他们还愿意从装满硝酸的烧杯中取出一枚银币,甚至愿意将硝酸泼向一名实验者,同样这些都在隐形玻璃的保护之下。然而,那些不容易受催眠却要求他们模拟被催眠组的参与者以及未被催眠的参与者也都毫无例外地同样答应实验者的要求。难道这些参与者对伤害他们自己和实验者如此漠然吗?当然不是,当实验者在实验之后与他们交流,他们说自己在实验中觉得相当安全。他们知道实验者不会让他们伤害到自己,安全是实验中的一个需求特征,而且参与者知道这一需求特征。在上述实验中,需要通过模拟控制组得到确认,是需求特征而不是催眠决定了最终的行为。

实验者—参与者的其他关系

本章开始讲的都是未参加过实验的参与者。实验者都喜欢这样的参与者。但并非所有没有经验的参与者都是好的,甚至大部分参与者并不是毫无经验。到目前为止,我们已经讨论了尽可能使参与者缺少经验的方法,或至少我们能觉察出参与者不是新手。但我们可以采用另一种替代的实验者—参与者关系。我们可以接受参与者不是毫无经验的事实,并且利用他们的问题解决能力。

欺骗和角色扮演

使用这种问题解决能力的一种方法是给参与者错误的线索,从而使他们对需求特征产生错误的解释。这种欺骗过程是为了隐瞒或伪装实验的真实目的。就道德和实际运用而言,欺骗在心理学中一直是一个有争论的话题。

欺骗在心理学中广泛被使用,尤其是在社会心理学的一些领域。事实上,如果不使用欺骗,社会心理学的一些领域就不能开展实验研究。例如,你对使旁观者帮助身处困境的人的条件感兴趣,让你站在街角去等待有人真的遇到了麻烦,这很困难。相反,你可能创造一种情形,让假冒者①陷入困境。然后,观察旁观者的行为。当然,你欺骗了旁观者,但是,除此之外你又能有什么更好的办法去开展这项实验呢? Stanley Milgram(1963)曾经做的一些著名的或臭名昭著的实验中都有欺骗。他使得参与者相信实验者对其他参与者做一些危险甚至是致命的电击,这是一个非常安全的认知心理学实验。例如,在一个偶然学习实验中,让参与者看一系列词;然后,在一些维度(如情感方面)对这些词进行等级评定。在实验结束后对参与者进

① 不,并不是一个叛乱士兵!在心理学中他们是指那些经过训练的人,在实验中按规定行事来帮助实验者。

行记忆测试,要求他们回忆那些词。参与者在某些方面已经受到了欺骗,因为他们并没有被告知他们要记忆这些词。但是如果提前告诉他们,那么,就无法研究无意学习,学习就可能是有目的的,而不是偶然的。

对于在实验中究竟是应该减少还是增加使用欺骗还存在争论。根据一些调查结果,在 20 世纪 70 年代到 80 年代,欺骗的使用的确增多了(Gross & Flemming,1982)。但是,现在的一些调查表明,使用欺骗趋于平稳甚至下降(Nicks, Korn,& Mainieri,1997)。但可以肯定的是使用的欺骗类型已经发生了变化,公然误导参与者的研究减少,更多的是仅仅保留了一些信息。

在研究中使用欺骗的基本逻辑是:尽管撒谎是不对的,但我们暂时误导参与者是合法的,因为我们是在为科学的发展作贡献。正如我们所讨论的那样,在心理学的一些领域,如果不使用欺骗就不可能回答一些重要的问题。同时,我们在实验结束后告知参与者,在欺骗这个问题上我们非常诚实,这样就消除了欺骗所带来的大部分影响。

而反对使用欺骗的研究者认为,如果你喜欢,你可以使用“误导”一词,但这是“说谎”的一种委婉说法。世界上有足够多的不诚实,但不能借着科学的名义来实施不诚实。有多少这种“科学上合法”的实验造成了科学的巨大飞跃? 没有很多! 我们可以设计出能够代替这种实验的其他实验,如让参与者完成角色扮演。试图通过在实验之后告诉参与者之前的欺骗行为的途径来消除欺骗的所有影响,这种观点过于幼稚。出于实际操作方面的考虑还有两个问题:欺骗会增加参与者将来的怀疑性,并减少对心理学家的信任,给心理学带来不好的名声。在心理学中使用欺骗不值得,应该除去它(Ortmann & Hertwig,1997)。

Kimmel 针对这两个问题开展的一项研究(Kimmel,1996)认为,参与者的怀疑对研究结果的影响可以忽略不计。另外,这一研究表明,被欺骗的参与者并没有觉得自己受到了研究者的愚弄而不满;此外,欺骗也没有使得参与者对心理学或科学的态度产生负面影响。例如,Christensen(1988)回顾了那些参与者对欺骗实验反应的研究发现,参加了欺骗实验的人报告他们并不介意受到欺骗,比那些参加非欺骗实验的人更喜欢实验,获得了更多的教育启发;他们并不认为自己的隐私受到了侵犯。此外,调查还表明,在一般人群中大部分人对于在实验中使用欺骗并没有强烈反对。

角色扮演被认为是欺骗的一种替代性过程。它是否同样有效? 一些实验者试图在相同的条件下使用欺骗和角色扮演;然后,比较这些结果。在角色扮演中,实验者要求参与者想象他们正处在一个特定的情境中,并作出他们认为在相似的真实情境中会作出的反应。例如,如果你对谈判行为感兴趣,你要一个人想象他是劳工领袖,另一个人想象她是一个公司主管,第三个人想象自己是仲裁人。你假设他们的反应与真实情境下人们的反应有些类似,然后开始实验。

不幸地是,尽管一些实验者报告欺骗和角色扮演可以产生相同结果(Greenberg,1967),但大多数人并不这么认为(Orne,1970)。此外,也很难描述在哪种情况下这两种方法可以得出相同结果。角色扮演在许多方面与前面提到的模拟控制非常相似。角色扮演可能仅反映了实验的需求特征,而并不能使我们预测在真实世界中会发生什么行为。

美国心理协会在它的《心理学家的伦理原则和行为准则》中提到了欺骗(APA,

2002)：

8.07 研究中的欺骗

(a)心理学家不能在研究中使用欺骗，除非实验者在使用欺骗装置时能给出充分理由，证明该研究的科学前景、教育或实用价值；并且没有欺骗性的其他替代过程可供使用，此时才可以使用欺骗。

(b)心理学家不能在有可能对参与者造成痛苦或者严重精神困扰的研究中欺骗参与者。

(c)心理学家应尽早向参与者解释，欺骗是实验设计和开展实验的一个整体特征，最好在他们确定参加实验时就告知，而不是等到完成数据收集之后才告知；并且允许参与者收回他们的数据。

作为一个初出茅庐的心理学家，你应该严肃地遵守这些规则。如果你想在实验中使用欺骗，你应该仔细权衡它的代价和收益。

自然观察

在第1章中已经提到另一种标准的实验者—参与者的关系。自然观察取决于实验者是否是一个显眼的观察者。例如，不是让参与者假装在扮演谈判角色，而是要求实验者去真实的谈判场所进行观察。实验者很少对环境的变量进行控制。他们通常需要耐心等待各种事件自然而然地发生；甚至不会控制潜在的混淆变量，或者不会从相关的数据中得出因果关系。

在这里我们把参与者当成是毫无经验的、单纯的观察者。至少我们应该意识到参与者的问题解决特性；然后设计实验，以便将影响他们尝试解决问题的效应计算出来。如有可能的话，这种尝试应该有利于我们，而不是与我们唱反调。

就以人类为对象的研究者的责任而言，我送上APA最后一句话。研究者应该很好遵守以下一些原则(APA,2002)：

1. 评价实验伦理的可接受性。
2. 判断参与者是否处于危险之中。
3. 保留伦理程序的责任。
4. 告知参与者风险，并获得知情同意书。
5. 判断欺骗是否合法且是否必不可少。
6. 尊重参与者拒绝参与实验的自由。
7. 避免参与者不适、受伤害和处于险境。
8. 实验后撰写任务报告。
9. 去除参与者任何令人不快的结果。
10. 为个体的研究信息保密。

在你遇到疑问去听取经验丰富的研究者建议的时候，如果你依据这些基本原则，你将永远不会在人类参与者的伦理方面出问题。也许最好的建议与你的态度有关。实验参与者通常都是在帮助我们。没有他们自愿参与实验，人类行为的科学研究会戛然而止。请对实验参与者始终抱着一颗感激之心。

公正对待动物

当我让心理学系的学生描绘出一个实验心理学家时，他们中的大多数人会将实验心理学家想象成一个穿着实验室的白大褂，在迷宫中赶老鼠的书呆子。这种描述有一点欺骗性，不仅仅是因为并非所有的心理学实验家都是书呆子，而且也因为只有7% ~8%的心理学研究涉及动物。而这些使用动物的实验90%使用的是大鼠、小鼠和鸟类，很少有人在实验中使用狗、猫或者非人类的灵长类动物。最近的一项调查表明，在大学心理课程中，62%的实验者使用动物的目的只是为了教学（Cunningham, personal communication, August 24, 2001）。最常用的活体动物是大鼠（81%）、鸟（27%）、小鼠（19%）和鱼（13%）。

为什么心理学要用动物研究

尽管在心理学研究中很少使用动物，但问题是为什么要用动物？为什么不全部以人类为研究对象？

行为的连续性

作为科学家，我们相信持进化论，并假设动物王国的连续性不仅仅是体现在生物性方面，而且也体现在行为上。虽然灵长类动物并不像人一样行为，而老鼠的表现更差一些；但人类的神经系统与动物的神经系统构成单元都是一样的，存在共同性。因为某些行为能力在进化史的早期阶段就出现，所以，很多人类的基本行为模式也出现在非人类动物当中。动物研究是基于这样的假设，即我们可以通过低等动物来研究某些普遍的基本行为。我们知道，进化中的动物保持了其基本行为，但发展出的复杂行为取代了基本的行为模式。因此，如果我们对研究基本行为模式有兴趣，如简单的学习或者动机，那么，使用动物不仅是可能的，而且也是应该优先考虑的方法，因为它可以避免基本行为模式被高等动物的智慧模式掩盖。但是，当我们试图将低等动物的行为推广到人类时，我们必须加倍小心。人类显然比大鼠要复杂得多，没有哪项权威调查证明大鼠的行为与人类的行为确切一致。尽管有某些不出名的动物研究因无知或头脑简单而过分夸大研究结果，但这种偶尔的误用并不能使动物实验的基本前提失效。

控　制

在研究中使用动物除了理论上的原因外，还有很多实际应用上的原因。其一，动物几乎可以随时得到。另外，在校大学生一定要过周末和放假，但动物可以长时间地供实验者使用。同样，无论是实验中还是实验之外，对动物所处的状态条件进行控制都是可行的、合理的。因此，动物实验可以研究一些有趣的变量，如过度拥挤、感观剥夺、睡醒周期和环境应激等。

我们可以控制动物的遗传和环境，通过动物的快速繁殖和多胞胎很容易实现。在人类研究中，遗传很少被当成控制变量或恒定变量。然而，在动物研究中，却可以这样做。但是，动物并不是万能的，做什么都用动物的观点也是错误的。我们将简

要地评价有关动物的伦理问题。

独特性

有些动物具有独一无二的特征,因此,它们更适合某一类型的研究。例如,果蝇不仅繁殖快,而且具有大量简单的染色体。乌贼拥有的神经元细胞数量比人多,因此,可以用来研究神经系统的结构。简言之,很多动物的中央神经系统都占有大量比例,以完成听觉或平衡感觉功能。在上述情况下,人类并不是最适合的研究参与者。

不可逆效应

最后,低等动物经常用于动物的功能或结构的不可逆效应的研究。切除研究只能用动物来做,因为它要求将神经系统的某一部分损毁以观察其行为的变化。同样,人类被试也不可能被用于需要在中枢神经系统植入电极的实验中。在很多情形下,这类研究也同样要求损毁动物部分脑区,并通过组织学①分析特定结构的变化。



例如,将动物进行社会隔离也会产生不可逆转的效应。一个著名的实验程序就是幼猴在出生后就立即与母猴分离,然后,将它与不同的人造母猴放在一起,以确定母性中哪个维度最重要(Harlow,1985)。人们不会使用婴儿做这类研究,并且有些人也同样不同意用年幼动物做这种实验。

动物伦理

不同的历史文化对于人类和动物的关系有着不同的定义。在基督教的传统中,人类有权支配动物,动物的出现就是为了人类专用的。这种观点在西方世界中被广为接受,并且一直持续到进化论之父查尔斯·达尔文时代。达尔文关于动物王国是一个连续体的命题是一把双刃剑。一方面,动物因其连续性变成了科学领域重要的研究对象。另一方面,这一连续性消除了人类高于其他动物的独特地位。在这样的时代背景之下,到19世纪末,第一个动物权力组织——反活体解剖组织——开始对人们的观念产生明显影响,尤其在英国(Dewsbury, 1999)。在20世纪开始的几十年中,反活体解剖组织不仅活跃于欧洲,而且在美国也很活跃。他们与很多著名心理学家,如威廉·詹姆士、巴甫洛夫、坎侬、约翰·华生等都有过口舌之战(那时以口头形式为主)。

此后,动物权利运动逐渐衰退,直到民权运动和Peter Singer(1976)的《动物解放》一书的出版之后,动物权利运动才又兴起。你可能也了解一些近代的动物权利运动。仅在美国大约就有7 000个动物保护组织(Justice Department,1993)。其中最大的组织是人道对待动物者(People for the Ethical Treatment of Animal, PETA),大约有350 000名参与者,职员70人,经费预算700万美元(Meyers,1990)。与活体

① 组织学通常是检验神经系统组织,以确定哪些组织被损毁以及电极安放的位置。

解剖组织的活动特点不同,现代运动组织是一种恐怖主义组织。武装性最强的组织是动物解放阵线(Animal Liberation Front, ALF),这个地下组织声称,对60%发生在实验室或实验者中的恐怖活动负责。它已经被FBI列为恐怖组织。尽管PETA否认与ALF有任何正式的联系,但它同意宣传ALF的活动,因为它相信,摧毁一些设施也是表达对虐待动物的愤怒的一种方法。

有人相信,公众对动物权利的关注已达到顶峰(Herzog, 1995)。1990年,在期刊综述上涉及这一内容的文章高达60篇,而在1994年只有大约25篇。然而,恐怖分子的活动仍在继续。1997年一个就职于某大学并在药物滥用方面获奖的研究者遭到学生组织的示威;该组织的一些成员和一个带着黑色面罩、自称是ALF成员的人出现在她家附近,对研究者及其家人进行骚扰,还威胁要焚毁她的家(APA, 1997)。据统计,动物权利组织发起了约313次恐怖活动,这些活动在1987—1988年达到峰值,随后渐渐衰减(Burd,1993)。这些事件和威胁的困扰耗费了研究机构数百万美元,并且逐渐威胁到研究者的生命(Mangan,1990)。为什么会这样?矛盾的根源是什么呢?

问题的一个方面是大多数极端分子认为,动物的权利和人的权利是一样的。因此,所有动物研究都应该被禁止。例如,一位拥护者认为:“目前,动物是最受压抑的群体。剥削动物与剥削黑人奴隶、童工和上世纪初期妇女堕落的罪一样重。这是我们这个时代的道德盲点。”(Ryder,1979:14)。主张废除动物研究的人包括PETA的发起人,她说:“为没有动物被关在牢笼之中而奋斗”(Havemann,1989)。普通大众、心理学家和心理学的学生是如何看待这一问题呢?

遗憾的是,还没有有关大众对心理学研究中使用动物的态度的调查。在医药领域,公众还是较为支持的。如果能获得关于人类健康问题的新成果,那么,可以允许科学家们开展可能引起疼痛和伤害的动物研究(如狗、大猩猩)吗?对这一问题的支持者从1985年的63%下降到1993年的53%(Pifer,Shimizu & Pifer, 1994, as cited in Plous,1996a)。反对意见在英国更为强烈。但在另一方面,一个民意测验发现,88%的人同意在药物实验中使用老鼠;相比之下,同意使用狗做研究的人只占55% (Associated Press,1985)。还有一些测验表明,超过四分之三的公众认为,“在医药研究中使用动物对于医学进步很有必要”(American Medical Association,1989, as cited in Plous,1996a)。

为了找出心理学家和心理学系学生对于在心理实验中使用动物的看法,Plous (1996a,1996b)进行了两个很好的调查。一个调查了3 982名心理学家,另一个调查了1 188名心理学专业的学生。令人吃惊的是,心理学家和学生对大多数问题的态度非常一致。当被问及是否支持在心理学研究中使用动物时,大约80%的心理学家和72%的学生表明他们强烈支持或者支持使用动物做研究。并且84%的心理学家和81%心理专业的学生认为,在心理学研究中使用动物对于心理学的进步很有必要。但事实上,使我惊讶的是47%的心理学家和44%的学生认为,他们不确定在心理学研究中使用动物是否人道。让我们看一看对待动物的准则。

很多年以来,美国农业部门都依据联邦章程监控大多数研究中动物(除了大鼠、小鼠和鸟)的使用情况。迫于动物权利组织的压力,1985年提出了动物权利法案的修正案,这一立法规范了如何对待狗、猫和非人类灵长动物。在进行了很多听

证会之后,1991 年通过了一系列的法规,为了遵从这些规定,研究机构花费了大约5.37 亿美元(Jaschik,1991)。这些法规关注的问题包括:对狗的训练、动物研究委员会的建立和对幼小动物的特别看护。尤其是对于灵长类动物的照看,应当确保它们身体和心灵处于良好状态①。

2002 年,经过了长期的关于实验室使用啮齿类动物和鸟类是否应有农业部门管理的争论,国会通过了养殖法案。国会认为,占非人类动物 95% 的大鼠、小鼠和鸟类等的使用已得到了美国国家健康研究院(NIH)、实验动物评估与认证协会以及和本地动物关爱与使用组织的重视。所有由联邦基金资助的研究机构在从事一项动物研究时,都要有一个由专家(包括一个兽医)和一名公众组成的动物研究委员会。所有使用动物的研究必须得到这个委员会的批准,并确保该研究遵守上述的规章制度。

关于这个主题,我要通过一些细节说服你:很多人包括心理学家都关心动物权力并且密切关注其中的变化。因为心理学家确实会用动物开展研究,所以,美国心理学会从来不会对这一问题保持沉默。心理学家的伦理法则和行为准则有一章专门介绍动物研究,我将在下面介绍这一内容。在 APA 的网站上有大量关于动物伦理方面的陈述,http://www.apa.org/science/anguide.html.

8.09 在研究中的动物关爱与使用

(a)心理学家获得、照看、使用和处置动物等应符合联邦、州及本地法律法规以及职业准则。

(b)受过研究方法训练和具有照看实验室动物经验的心理学家应该监督涉及动物实验的整个过程,并负责确保动物舒适、健康和受到人道对待。

(c)心理学家应确保所有使用动物的人接受研究方法的指导,并使他们能够照看、维护和处理所使用的动物,胜任自己的工作。

(d)心理学家应努力去减少动物的不适、感染、生病和疼痛。

(e)心理学家只有在别无选择的情况下和实验目的被证明是科学的、具有教育应用价值时,才能对动物实施疼痛、压力或使之处于被剥夺状态。

(f)心理学家在适当的麻醉之下进行外科手术,并且在术后避免感染和降低疼痛。

(g)心理学家应快速地、使痛苦尽可能小的方式终结动物的生命,处理过程应符合相应程序。

为什么心理学家们如此关注他们从事动物研究的权利可能会被剥夺?除了大多数与动物一起工作的心理学家喜欢从事这项工作,并且认为自己是动物爱好者外,他们还列举了开展这种研究以提高人类和动物的福利的进展(Miller,1985)。药品在使用前必须进行这类研究,动物研究帮助消灭了小儿麻痹症、狂犬病、霍乱和白喉等病症,也促成了胰岛素的合成和白内障手术的发展,治愈了很多淋巴性白血病

① 具有讽刺意味的是,正是受到动物权益保护激进分子强烈批评的研究者之一 Harry Harlow 做了很多研究,提出了这些规则。

的孩子。在心理学中,动物研究促进了精神紊乱的治疗、疼痛控制、药物滥用以及中风恢复等方面。最成功的行为疗法,也是在50年前的动物研究中发现的。在有些情况下,特别是在基础研究中,未来收益通常都是非常难以预料的。但是,大多数研究者认为,动物所付出的代价可以被潜在的收益抵消。

甚至大多数动物保护者也处于一种中性态度,他们希望通过一个聪明而理智的方法来解决这一争论。大多数动物保护者也谴责那些恐吓研究者的极端分子。如果你希望了解更多有关这一主题的内容,APA出版了大量这方面的守则,如《关爱和使用动物伦理准则》(1986)。Boyce(1989)在美国兽医协会期刊上发表了平衡处理动物研究的观点,Segal(1982)对科学与生命之间的权衡问题进行了很好的权释。Carroll & Overmier(2001)也出版了一本书《动物研究和人类健康:通过行为科学造福人类》,对这一问题进行了全面深入地探讨。最后,新墨西哥大学一位半身不遂的心理学家Dennis Feeney(1987)在《美国心理学家》杂志上对动物研究发表了强烈的宣言。他认为,对动物权利和人类权利的探讨主要来自于科学、农业以及动物福利组织中的人,但是忽视了残障人士的看法。他指出,在生物反馈和中风康复中的基本行为研究给治疗带来了无法预计的进步。

以下是他的立场的实质:

> 我们这些患有不可治愈疾病和永久性损伤的人只能希望通过研究进步来治愈。很多这类实验需要使用动物,因此,我们必须找到一种折中的办法共同捍卫人类权利和动物福利。在决定减少人类苦难和违反动物福利之间如何妥协之中,就包含了痛苦和不适,我无疑选择减少人类痛苦。(593页)

在最后,你将会决定你自己对使用动物研究的伦理问题。在作决定时,你必须了解自己对动物和人类关系的最基本信仰。如果你像大多数人一样,你会发现,人们的观点很难完全一致。例如,你吃汉堡吗?你同意将你的狗或者其他宠物用于医学实验吗?你能容忍牺牲100条狗来挽救你孩子生命的治愈方法吗?你买蚂蚁和蟑螂喷雾剂吗?你愿意收养本地收容所中所有的猫吗?从你对这些问题的回答中,你能决定自己的立场并产生与之一致的人生观吗?

这不是一个非此即彼的问题。正如大多数人一样,你可能简单地权衡每一种情况,并试图判断是否利大于弊。就本质而言,有些心理学实验使动物受到压力和疼痛。在这种情形下,你应该在实验开始前就说服自己潜在的科学上的益处会大于损失,并且捍卫自己的决定。尽管研究机构确实通过专家委员会筛选动物研究,但你应该看到专家委员会所用的标准只是最低标准。你应该符合更严厉的伦理标准——你自己的标准。

小 结

在本章中,我们讨论了公正对待动物和人类参与者的伦理和方法方面的问题。因为实验参与者的权利主要取决于实验者,所以,实验者要遵守一些基本的礼貌规则。他们必须在场、准时、充分准备、有礼貌、保密及有专业素养。机构的审查委员会对研究报告进行筛选,并帮助研究者以符合伦理道德的方式对待动物和人类参与

者。在这些问题中,他们关心的是这些参与者在参与之前是否获得了知情承诺。

尽管过去我们曾经假设人类参与者是天真的、被动的,但他们都是现实问题的解决者,他们对实验情境中的需求特征或隐藏的线索都非常敏感。他们如何对待这些需求特性取决于他们在实验中是否合作、防御或不合作。我们可以通过实验程序自动化、单盲或双盲实验设计最大程度地减少这些需求特性(使参与者或参与者和实验者都不清楚具体操纵的条件),也可以使用多位实验者。在毒品研究中,一种双盲设计使用无疗效的安慰剂作为控制条件。这种情况下使被试不了解实验操纵是非常重要的。实验者必须知道可能会出现霍桑效应,即行为变化仅仅归因于给予了被试关注。如果没有需求特性在实验中出现,我们有时可以通过实验后的问卷、非实验或者模拟控制组来甄别它们。另一个假设是被试是天真的,利用他们解决问题的本质特性,给以虚假的需求特性,使他们无法了解实验的真实目的。替代欺骗的方法是要求被试进行角色扮演或者在自然情境中观察他们的行为。

心理学家也使用动物做实验,因为动物的一些基本行为方式与人类行为具有连续性;而人类行为更加复杂,容易产生混淆。动物也为环境和基因控制提供可能性。有些动物拥有一些独特的特征,这使得它们在某些类型的实验中占优势。动物权利运动的历史很长。在过去的几十年中,这个运动得到了一些极端分子和暴力拥护者的支持。公众、心理学家以及心理学系的学生都赞同动物研究。近几年中,在联邦、州和研究机构中关于动物研究的规范、法律等已经制定。所以,虐待动物的可能性变小。很多研究者都相信动物研究利大于弊,并且认为很多心理学中的动物研究成果都能够造福于人类,如治疗精神异常和滥用药物、疼痛控制和中风康复等。

5

如何公正地对待科学

科学就意味着接受事实，即使这个事实与愿望相违背。

B. F. SKINNER (1953)

为了获得某个结果，人必然期望获得这一特定结果：如果你想要得到一个特定的结果，你就会获得它。

T. D. LYSENKO, QUOTED IN I. M. LERNER (1968)

科学中的欺骗不仅仅是桶里的苹果烂了的问题，它还与桶有关。

NICHOLAS WADE, QUOTED IN K. MCDONALD (1983)

研究是一项团体性活动，需要参加者相信同事的正直。

ARNOLD S. RELMAN, QUOTED IN K. MCDONALD (1983)

在我们的文化中总会出现不诚实行为，然而，最令人称奇的是，在所有文化中它都被当做是人类的主要恶习。我们应当看到，这种知觉总是与对谎言构成的疑虑相一致，更与在诸多谎言得逞情境下的暗自怀疑相一致。

TONY COADY（澳大利亚心理学家）

道德地活着远远好于不道德地死去。

MARTIN E. P. SELIGMAN（美国心理学会主席，1998）

本章将继续讨论研究道德问题,但在这里我们要讨论的是公正地对待科学。在某些方面,科学的自身防护功能要比一个研究参与者弱。动物受到虐待时会挣扎、惨叫,甚至死亡。人受到虐待时也会挣扎、惨叫,甚至会去上诉。而科学连挣扎和惨叫都不会。但是,假如你长时间虐待它,你的科学家同行最终可能会挣扎和惨叫。

你可能有些疑惑不解,自己如何能对无生命的科学有不公正呢。在一定意义上,科学能够被看成是有生命的,因为它是不断运动变化着的,并且我们希望它是不断扩展的知识体系。新的研究不断替代或在已有理论基础上构建新的知识体系。在科学上,任何阻碍科学的发展或者导致它误入歧途的行为都是不道德的。

科学中建立了一些自我防护措施,它可以保证科学知识体系沿着正确的方向不断发展。例如,你的研究成果在科学刊物上发表前,一组有所成就的科学家会被选作评委对它进行评审。通过评审,评审者将判断研究是否符合实验规则,本书已对这些规则进行了讨论。评审者还将确定你的成果是否有足够的理由来占用杂志有限的版面。评审者和编辑通过对研究的筛选,使优秀的相关研究发现纳入现有知识体系之中①。尽管这个评审过程不完美,但大多数的心理学家相信它完成了最重要的筛选功能。

设计评审体系是为了排除拙劣的或缺少足够成果的研究,而不是为了评判那些有能力做出优秀研究的研究者是否在他们的研究中撒谎。科学家本人应该知道这些规则,并用实际行动保证自己做到了这一点。然而,这一行为经常受到个人追求功利思想的挑战,例如,“除非我发表了 5 篇文章,否则,他们不会提升我”“我必须使自己的研究结果看起来好些”;或者“我必须有一个好的实验结果以便能够晋升”。

因为不道德的行为会严重影响我们的科学,而且我们也没有很好的预防措施去检测它们,所以,当这种行为被发现后,科学界是无法容忍的;这样的研究者也将失去进行科学研究的基本权利。这些无法接受的不道德行为被称做肮脏伎俩(dirty tricks),在本章第 1 节将进行讨论。其他节将讨论一些需要防范但并不是极其恶劣的行为,这些行为不至于使一个研究者失去研究的基本权利,我称之为可疑伎俩(questionable tricks)。最后,在科学中有这样的情况,一个研究者报告了一些真相,尽管并非是全部真相,但科学从这些行为中是获益而非受损;这种行为可以使我们的科学更加有效和更加易懂。因此,这些行为不仅是可以接受的,而且也是必须的;我称这些行为是简化伎俩(neat tricks)。

肮脏伎俩

造　假

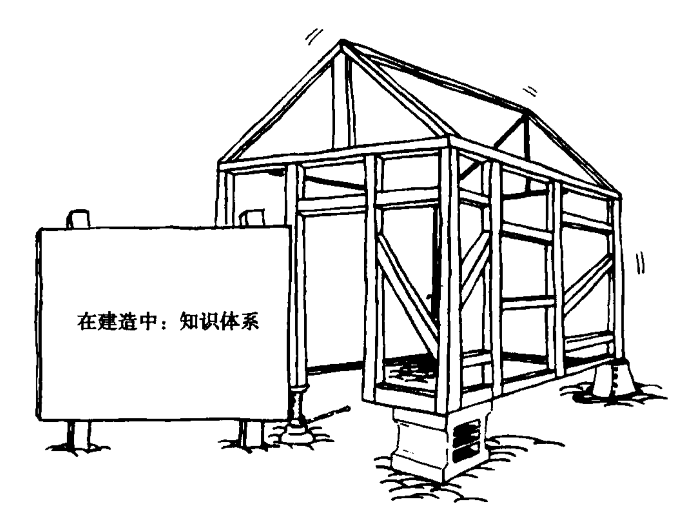
一种公开的欺骗就是数据造假。一些“实验者”知道做实验最简单的方法就是根本不去做。他们不需要为一些烦琐的事情而烦恼,如购买仪器、招募被试或学着

① 注意,这里的相关一词的含义与人们常用的含义不同。这里指的是与科学相关,而不是与流行的话题有关。有时候,与后者相关的主题却恰恰与前者相关最小。

做统计分析。他们所要做的只是学会写实验报告。(造假者应该仔细阅读第 13 章。他们应该学习另一个专业,因为他们最终不会成为心理学家。)

作为一个学生你可能会尝试伪造数据,如一个作业提交期限快到了,而你还没有完成它。千万别那么做!晚交作业得到的分数可能比较低,但伪造结果将导致退学或得到一封很差的推荐信。专业的科学家是完全不能容忍这种行为的。

美国心理学会(APA)颁布了一个"心理学家道德准则和行为规范",它试图明确心理学家在研究中应遵循的道德准则。最新的标准颁布于 2002 年,它说明了心理学的科学属性(APA, 2002)。根据标准 8.10,"(a)心理学家不能伪造数据""(b)如果心理学家发现了自己已发表数据中存在明显错误,那么,他们必须对这些错误进行更正、撤销或采用其他恰当的途径进行公布与说明。"



让我们回到 20 世纪早期,一位叫 Kammerer 的生物学家试图证明后天获得的特性可以遗传(Ley,1995),这个观点与达尔文的进化论完全不同。他声称已经在黑土中让火蜥蜴延续后代。他不但报告说黑底黄点的火蜥蜴身上的黄点随一代代地繁衍而变小,而且这些点大小的减少也会遗传给下一代。另一位研究者对此观点提出了质疑,并考察了 Kammerer 所报告的繁殖代数所需要的时间;该研究者发现,所需时间比 Kammerer 研究的要长得多。其他科学家也开始寻找其中的原因,直到 7 年后两位著名的科学家重新检验了一些样本。他们发现,样本中被注入了印度黑墨水。Kammerer 承认他的结果"是用印度墨水对死后的样本进行标注而变成的"。后来他因此而自杀了。

在科学中最卑鄙的行为是向知识体系中添加干扰。如果你做了一个很烂的研究,人们能够并将忽略它;但是,如果你没有做出很好的研究,而你假装自己做出了,那么,你将会阻碍知识体系的发展。其他人可能会延续你的研究或试图在你的研究基础上来进行他们自己的研究。直到最终才发现你的研究中存在错误。他们可能已花费了很多时间去关注这些结果,并且可能需要重新解释所有的结果。这种欺骗被发现得越晚,最终浪费的科学资源就越多。

如果这种行为没有及时被发现,那么,对科学的欺骗可以导致对社会的欺骗。例如,在 20 世纪 30 年代后期,一位叫 T. D. Lysenko 的俄国人认为,获得的特性可以被遗传(Lerner, 1968)。为了使自己的理论更具说服力,他伪造了大量的数据。

Lysenko 们声称,他们发现了奇迹般的结果,如将麦子转变为大麦、裸麦、燕麦,甚至矢车菊;甜菜转变为大白菜;松树变冷杉;角树丛转变为胡桃林(用伪造的照片作为证据),甚至能够用莺鸟下的蛋孵出杜鹃来(Lerner, 1968)。

Lysenko 不道德的行为不仅损害了科学,而且对社会造成了不良后果。他应该为俄国大量的遗传学家被免职、流放和处决承担责任。他也曾使斯大林和赫鲁晓夫认为,他的理论是正确的,可以广泛地应用于农业项目中去(Medvedev, 1969)。Lysenko 的方法错误导致农业失败,Lysenko 失败了,赫鲁晓夫失败了(尽管并非仅此原因),俄国社会遭殃了。

当我在本书的第一版中写这一章的时候,我必须回顾历史来寻找科学骗子的案例。不幸的是,在当今研究中更容易找到骗子的案例。我在一些专业的或地方的报纸上发现如下的标题:"指控剽窃科学手稿增加了对'知识剽窃'的关注""诺贝尔奖得主面临不端行为的指控"和"美国引入诉讼程序,指控那些有欺骗行为的科学家"。

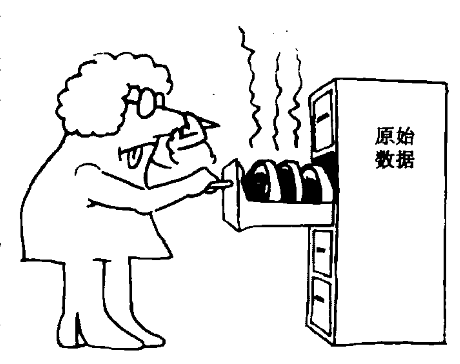
肮脏伎俩是新事物,还是因为现在人们关注更多才揭示了它们? 显然,即便是早期的科学家也不完全是干净的。19 世纪的遗传学家孟德尔(Gregor Mendel)被指控有一定程度的不真实(Fisher, 1936)。甚至牛顿(Isaac Newton)也曾报告过明显超乎他自己的测量能力的精确数据(Westfall,1973)。

最近,与心理学关系更密切的例子是,一个著名的超感知(ESP)研究者发现,他的实验室主任伪造了数据。一个富有怀疑精神的研究助理潜伏在他们的实验室里,发现了主任通过修改数据来支持超感知的结果。

在心理学中,最出名的例子是西瑞尔·伯特(Cyril Burt)爵士,一位英国著名的心理学家,他曾被乔治(King George)六世封为爵士。他提出智商(IQ)在很大程度上是天生的观点,而他有关双胞胎的研究正是支持他这一观点的主要依据。他去世之后,研究者发现,他报告的同卵双胞胎之间的相关系数精确到小数点后三位数并保持多年不变,而这期间数据中又增加了新的双胞胎样本。这看上去不像是巧合。在伦敦的《星期日泰晤士报》(*Sunday Times*)报道了造假的新闻,争论就此展开。一方的研究者认为,伯特的数据有假,甚至是伪造的(McAskie,1978)。另一方则认为,造假是不可能的,更有可能的是简单的疏忽(Jensen,1978)。不幸的是,这场争论被精英优越论和平等论的政治争论推向高潮。

我们如何防止研究者可能出现的不诚实呢? 怕是永远也无法建立一个简单有效的系统,因为建立这种系统所需的花费以及知识自由的丧失将摧毁科学的大厦。当前已经采取了许多预防措施。美国国立卫生研究院(National Institutes of Health)下的公共卫生署(Public Health Service)已经要求大学必须颁布避免不端行为的准则(Cordes, 1990)。美国大学联合会建议公共机构应该为相关标准建立明确的政策,以便管理者执行这些政策("AAU Statement", 1983)。当一个研究者得到联邦政府资金的支持后,我们甚至应该有一个"揭发者"法律,这种法律其实早在 19 世纪就有了,它允许他人对研究者、大学提出诉讼,并获得适当比例的政府资金的奖励。最近,一个研究实验室里的前技术人员对一个研究者及其大学发起了诉讼,起诉这位研究者伪造数据(Cordes, 1990)。

一种减少伪造的方法是为研究的原始数据建立档案(Bryant & Wortman, 1978)①。调查者在试图证实伯特的研究时遇到的一个最明显的问题是他的数据储存非常草率。有关双胞胎研究的初始测验结果被塞入6个茶柜里,后来它们全部被损坏!通过数据档案,研究者可以将数据独立存储起来,或把数据存放在一个存储中心。通过这种方法,大众就可以获得研究数据。但是,情况往往并非如此。例如,一个研究生向37位文章作者索要原始数据,结果只有24%的研究者同意提供(Wolins, 1962)。



存档还有以下的好处:研究者可以对他们的原始数据进行更加仔细地分析;通过检验事先没有确定的问题,丰富知识体系;可以比较不同年份的数据,从而进行纵向研究。

建立正规的数据存储系统需要一定代价,它包括设置管理机构、复制数据或使之标准化花费的时间和金钱,还有对正直科学家信任的丧失。有人认为,这种解决方法代价太大,而解决的问题相对较小。他们相信反复的检查与比较就足以发现很多造假行为,从而使肮脏的伎俩变成罕见的而不是常有的事情。

无论是否有正式的要求,你至少应当将数据保留5年。研究者常常需要从别人那里获得数据。随着电脑的使用,原始数据的存储和提取也变得非常简单。事实上,APA的另一个道德标准是8.14:"(a)研究结果公布后,假如除参加者的隐私需要得到保护或与法律相关的专利数据不能公布外,心理学家不能拒绝将自己的原始数据给那些试图进行验证性的重新分析的同行或是仅仅为了分析这些数据的人。"如果你形成了存储原始数据的习惯,将有助于保护科学;同时,也保护了你自己,避免受到错误地指控。

有些数据确实表明,科学研究中的骗子相对较少。例如,从1982到1988年,在美国国立卫生研究院(NIH)资助了50 000名科学家的研究中,仅处理了15到20项不端行为的指控。另一方面,据一项对某研究型重点大学的科学家的调查,只有三分之一的人怀疑过他们的同事剽窃或伪造了数据,但是,其中又只有不到一半的人会采取措施,求证或报告他们的怀疑(Hostetler, 1988)。因此,如果科学家不能采取更加正式的程序去发现和减少科学上的不道德行为,那么,国会将采取一些法律行动。但这可能又会给很多科学家带来困扰,因为他们相信,由这些未经训练、缺乏科研经历的人来制定规则会对科学造成不可预想的后果。

防治科学研究中的造假行为的最好和最有效的方法是向研究新手强调道德行为的重要性。这正是你读这一节的目的。只有理解科学工作的方法,才能树立道德行为的真正动机。除非我们相信同事的研究结果,否则,我们也都无法继续构建科学知识体系。

① 在将因变量综合起来进行统计分析之前的对因变量的单个测量是你的原始数据。与肉不同,原始数据不会腐烂——只会占用研究空间。

剽 窃

APA 道德标准的 8.11 条对剽窃进行了说明:"即使别人的文章或数据经常被引用,心理学家也不能将其他人的文章或数据据为己有。"这一陈述看起来很清晰,但最近几年剽窃已成为人们关注的焦点,因此,很有必要对它进行详细说明。实际上,在过去 20 年中,很多大学校长因为剽窃讲稿而丢了工作,很多作者因他们的著作中存在大部分剽窃他人成果的内容而受到指控。这些行为并不意味着我们所说的剽窃标准有所改变,而是大众逐渐忽略或轻视了这个标准。

大多数人知道,直接引用其他人发表的研究内容即为剽窃。然而,很多学生相信互联网上的材料是可以自由猎取的猎物,可以将其归为己有。我要强调的是,盗用互联网上的内容与盗用传统的正式出版的资源同样恶劣。教授们可能会使用如 http://www.turnitin.com 来识别抄袭互联网内容的学生。



我所见到的最坏的欺骗个案是一个咨询心理学家,在他 120 页的博士论文中,有 90 页是从两本书上抄来的。尽管他在参考文献中列出了这两本书,但在原稿中并没有任何正式的引用,而是逐字的抄袭! 然而,这是一个看上去很好判断的个案,这个学生的前导师宣称,他不认为该学生的行为是剽窃,因为该学生不是故意去剽窃的。我再次重申:意图不重要,行为决定了它的性质。这个个案中的心理学家丢掉了他的博士学位和他进行专业研究的权利。

剽窃不仅包括摘取别人的语句,盗窃别人的观点和盗窃语句具有同样的性质。就像第 13 章中所说的,心理学家和其他作家不同,他们不能在报告中经常直接引用别人观点。我们更可能是意译其他研究者的观点,也就是用我们自己的语言重新组织一个特定的观点或想法。但是,意译不能将我们从正当引用的束缚中解脱。当描述其他研究者的想法、理论或猜想时,你必须标明这些研究者或这些观点的来源。

为了说明我们所讨论的剽窃的几个类型,假设你读了一篇由 Diener, Lucas 和 Scollon 发表在 *American Psychologist*(2006)的一段话:

> 享乐适应理论建立在一个无意识的习惯化模型上。在该模型中,心理系统对偏离个体当前的适应水平作出反应(Helson, 1948, 1964)。无意识的习惯化

过程具有适应性,这是由于他们可以使经常出现的刺激在背景中淡化。因此,仍然有资源处理新刺激,这些新刺激往往需要迅速的注意(Fredrick & Lowenstein, 1999)。

在报告中你会这样写吗?

享乐适应理论建立在一个无意识的习惯化模型上,在该模型中,心理系统对偏离个体当前的适应水平作出反应。无意识的习惯化过程具有适应性,因为它们会使恒定的刺激消失在背景中。

不,你不能这样写,因为你用了他人的话语,而没有标明出处。那么,你能这样说吗?

享乐适应理论的基础是无意识习惯化模型。由于存在对某一特定习惯化的适应水平的偏离,从而使恒定的刺激变成了背景噪音,使得资源可以被用于加工新信息。

尽管你已经对话语进行了释义,但你仍然是在说别人的观点,你必须给出这些观点的来源。那么,你也许可以这样说?

享乐适应理论的基础是无意识的习惯化模型(Helson, 1948, 1964)。由于存在对某一特定习惯化的适应水平的偏离,从而使恒定的刺激变成了背景噪音,因而使得资源可以被用来加工新信息(Fredrick & Lowenstein, 1999)。

这段话比较好,因为它给出了观点的来源。但是,如果你只是读了 Diener, Lucas 和 Scollon 的文章,你就不能这样引用。通过列举 Helson, Fredrick 和 Lowenstein 的文献来源,说明你阅读了这些文章并且进行了释义。但是,你所知道的来源只是来自 Diener 等人的文章,那么,其他的两个文献来源被称为第二来源(secondary sources),你只能在主要来源中进行引用。例如,"Helson(1948, 1964),正如 Diener, Lucas 和 Scollon(2006)所引用的那样,一个无意识的习惯化模型是……"。若非确实没其他办法,最好不要引用第二来源的文献,你应该找到第一来源的文章,然后,作为主要来源来引用。

我曾用了相当大的篇幅讨论了抄袭问题,因为这是一个相当严重的违反道德的问题。甚至很小的错误都会导致相当严重的后果。严重的错误将会导致名誉扫地、失去工作,与例子中的学位论文事件一样,丢掉博士学位。

篡改资历

如果你走到朋友身边对他说:"你能够为我做倒立吗?"他的反应可能是"为什么?"然而,如果你走到另一位朋友身边说:"我正在做一个实验,你能够为我倒立吗?"你朋友的反应很可能是"多久?"这种差异是因为社会给予科学家很多特权而普通民众却没有这种特权。我们允许科学家,尤其是行为科学家,自由地开展实验。因为就社会而言,我们觉得我们从这些实验中所得到的经常要比付出的多。我们经常会给予科学家一定的威望,一般人都会顺从他们的要求。

科学家们不仅仅被允许用合适的方法去操纵他们周围的人,而且他们有时候还可以得到国家税收的支持。然而,当我们相信他们对社会的益处不再高于他们的害

处时,我们有能力取消他们的这些特权。要求取得专业的实验资格证书就是我们的一项政策。因此,要向其他科学家证明你是一个有资格的研究者,你必须向他们展示你的专业素养,一般以个人简历或履历的形式展示。履历中的记录体现出你的专业性,其中包含教育程度、工作经验以及出版的书籍与文章。你用它可以进入研究院,获得专业认可,或者是找到工作。毫无疑问,这些文件必须全部是准确无误。无论如何,篡改资历都是无耻肮脏的伎俩。

APA 在这个问题上有很明确的道德标准。标准5.01 防止虚假或欺骗的条款提供了下面的指导细则:

(b)心理学家在以下方面不得造假、欺骗或不诚实:

(1)他们的训练、经验或能力;

(2)他们的专业学位;

(3)他们的证书;

(4)他们的学院或隶属的协会;

(5)他们的任职;

(6)科学或客观的偏向或他们任职的成功程度或结果;

(7)他们的薪水;

(8)他们的出版物或研究成果。

在我职业生涯的早期,我见过这样一个事件:一名天资聪明的学生使用伪造的履历试图进入研究院。他有良好的记录和一堆教他的教授们给他写的推荐信,但他的履历中有几篇论文并不存在。当他的教授发现了他的欺骗行为,这名学生就失去了良好的记录和推荐信,也不可能成为今天的实验心理学家。因为科学家和社会之间的协议非常脆弱,这种不诚实打破了两者之间的微妙平衡,是不能容忍的。

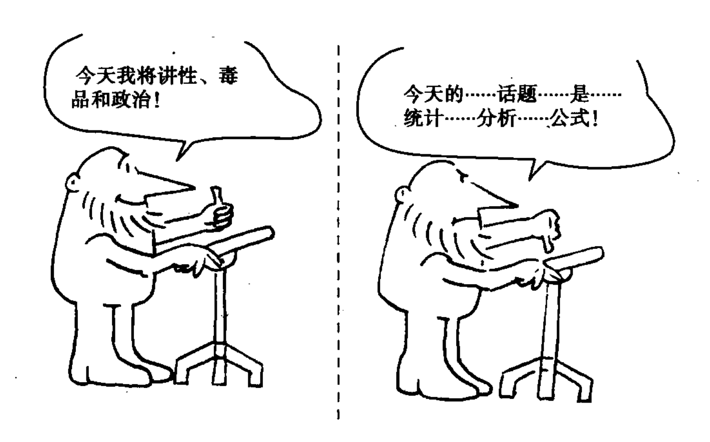
我希望我们讨论这种肮脏的伎俩只是在浪费时间,我们可能并没有想过要做这种事情。但我相信这些问题必须在实验者受训早期就要牢记。做一名心理学实验者会非常有趣,但实验的真实目的是构建科学。那些不想遵守用来保证这个过程正常运转的规则的人应该被排除在科学之外。

可疑伎俩

大多数研究者不能接受,并遭到反对和责难甚至是唾弃的行为被视为可疑的伎俩。这些行为可能会发生在实验设计、实验过程、数据分析或实验报告等过程中。

实验设计

在前面的章节中,我们通过需求特征讨论了实验者偏向。在设计实验中,需求特征可能在因变量中引起需求变化;如果你不试图最大程度减小需求特征或者甚至不去发现它们,那么,你是在有意识的不诚实。如果你将一个特殊的变量作为控制变量,而事实上这个变量和自变量之间存在系统变化关系,那么,你的实验可能产生了混淆。在一些非实验室实验中这些混淆物非常难控制,但在很多情况下我们有理由称之为欺骗。



例如,在第 2 章中讨论的一个实验,我试图用不同的演讲速度给初级心理班做讲座,以确定不同速度是否可以决定学生注意力。在一些天我用很慢的速度,另一些天用中等速度,其他时间速度很快。我们通过记录背景噪音的等级来测量注意力。这类实验很容易产生偏向。我改变的可能不仅仅是讲课速度,还可能包括我所讲话题的生动程度或过去我常用来支持我观点的有趣例子的数量。这些维度的变化比讲课速度更容易使自变量混淆,不管是有意还是无意的。

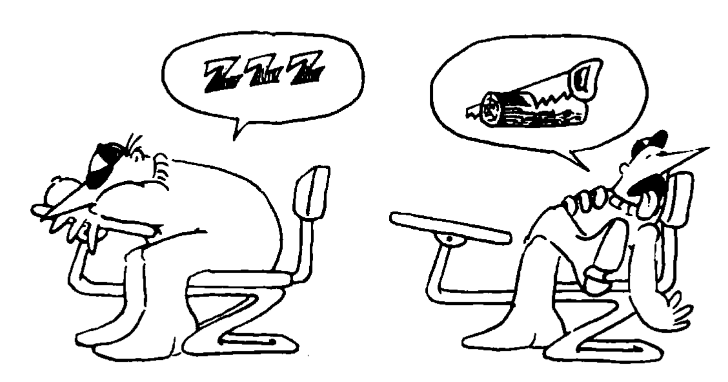
有一种方法可以最大程度减少这种虚假信息的可能性,即在实验设计上下工夫,让与结果没有利害关系的同事对每个讲座中可能混淆的变量进行评定。实验者只从那些评定等级相等的讲座中收集资料。当然,实验者偏向作为某种程度的欺骗并不一定出现在第一种设计中;但是第二种设计会更具说服力,因为偏向很难在其中产生。

数据收集

你在收集数据过程中也可能会出现不诚实行为,尤其在必须通过主观方式来判断被试的反应时更容易出现。例如,在以上讨论的实验中,假如实验者试图通过记录学生注意和不注意听讲的时间百分比,将学生的行为分为注意和不注意两类。假如一个学生坐在座位上用铅笔在纸上乱画。她是在记笔记还是在乱画呢? 另一个学生闭着眼睛坐在那里。他是在思考还是在睡觉呢? 我们可以根据自己的偏向判断不同行为。如果带有这类偏向的实验者做这种分类的话,它产生的问题是显而易见的。

为了避免这类虚假信息,实验者应该建立一个注意和不注意行为的标准核对表,同时,严格判断观察到的类型,并对学生行为进行独立归类。在可能的情况下,甚至让不了解实验目的人用录音机记录被观察的内容。这些预防措施降低了有意或无意产生的虚假信息的可能性。

偏向有时候也会发生在那些看似很公正地测量反应实验中。有一个实验要求被试移动一个小棒,将标记与一个移动目标相对齐。在每间隔 10 秒后,实验者快速地读出刻度盘上指针的读数并将其复位(离目标越远的标记,指针在刻度盘上的移动越快)。这个任务很难,因为指针在量度线上很少是垂直的。实验者关于刻度盘上指针位置的判断很可能会出现偏差。在这个实验中实验者要读刻度盘超过



哪个学生是集中注意力的?

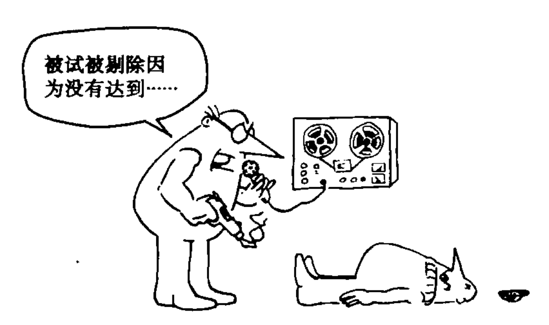
15 000次,这就增加了出错的可能性。在读数中即使很小的不一致都可能导致最终结果的主观偏向。因此,不论何时,存在偏向的实验者必须通过判断去解释每一次反应,他们还必须精心设计整个过程,以保证判断更为准确。

数据分析

在数据分析中你还应当避免使用有偏差的方法分析数据。正如第 12 章中讨论的那样,统计检验经常被用来决定某种结果是真实的现象还是由于偶然才发生。这些检验仅能用于特定假设前提下进行推断。当假设在总体上与事实相冲突时,检验就可能存在问题。

例如,常用的统计检验都要求数据分布接近正态。尽管不完全满足这个假设并不会使得检验失效,因此,当分布并不是正态时,有些研究者仍然使用这些检验方法。就一个实验而言,你首先要知道统计检验需要满足怎样的前提假设。如果不能满足一个或更多的前提假设,那么,检验就很可能会出错。

在你进行数据分析时,你可能会发现,尽管大多数被试的反应表现出预期的实验效应,但仍有些被试没有这些效应。就这一点而言,你对这些没有效应的数据无能为力①。显然,如果你把那些不符合预期结果的数据剔除,那么,你的实验会永远支持你的假设!正因为如此,你在根据因变量剔除被试数据时要非常小心。你绝对不能以因变量在自变量不同水平上的差异为依据剔除那些与自己的预期不一致的数据。



只有当被试的反应不符合你在实验前设定的标准,同时能够合理说明并把它详

① 除非你对研究个体差异非常感兴趣。

细写在实验报告中时，你才可以剔除这些被试。例如，假如你对噪音影响人们打字效率感兴趣。在实验开始前，你可能会将那些在无噪音背景中每分钟打字低于10个的被试从所有被试中剔除。你的理由可能是，这些人本来打字能力就很差，即使噪音对打字有影响，在他们身上也看不出这种效应。或者你可能说我只对噪音是否影响有经验的打字员感兴趣，而那些每分钟打字低于10个的人不是有经验的打字员。然而，如果你不对那些根据实验前预定的因变量水平剔除被试作出合理说明，那么，你就不应该去剔除他们。

根据某一标准而不是因变量的数值剔除被试数据更为安全。因此，在实验前就应该制定出剔除数据的标准，并在实验报告中对这些标准进行详细说明。例如，你要求被试看一排字母，然后，报告出红颜色字母。在这种情况下，你必须在实验前就排除那些色盲被试。

报告结果

假设你已经完成了对实验结果的分析，而且现在准备报告研究结果。你通常会想把结果用图表的方式描述出来。我们将在第12章讨论绘制图表的一些规则。有人曾写过一本有关如何通过篡改图表和统计结果来说谎的书（Best,2001；Campbell,1974；Huff,1954；Wainer,2000）。例如，有实验者可能会放大图的一个坐标轴，让微小的差异显得差异巨大，或者扭曲一个坐标轴的标尺，从而改变图的形状。如果你是一个富于创造力的人，你会找到很多方法让低质量的结果变得看起来非常漂亮。显然，这些行为不能推进科学发展，并被认为是不恰当的。

另一种形式的可疑伎俩是采用支离破碎的方法报告实验结果。尽管在研究中每次只能做一个实验，但你却不能用这种方式报告研究结果。几十年前，典型的心理学期刊文章都只报告单个实验结果。然而，近些年来，这个领域发展非常迅速，文献数量激增。如此多的人做了如此多的实验，以至于人们无法跟得上实验发展的潮流。因此，除非单个实验的研究报告对该领域的发展有着非同寻常的巨大贡献，否则，有些期刊不会接受它们。

通常情况下，研究报告应该包括从一系列实验中整合出来的研究结果。通过这种方法，可以使新发现变得更加富有成效、条理清晰；读者可以省去重新寻找自己研究定位、重新阅读每个实验的引言和研究程序以及将孤立的研究整合到一个相互联系的结构中去的时间。在当今这个“出版或退稿”的世界中，研究者为了增加自己的成果数量，可能会经不起诱惑而采用拆解的手段发表文章。然而，这些行为既不能提高研究者的声誉，也不能促进科学知识的发展。

简化伎俩

尽管看起来违反直觉，但有时候对阅读者的必要“撒谎”可以提高交流效率。研究常常是一个松散的过程。然而，当阅读实验报告时，你会觉得研究者的研究过程总是系统而秩序井然①。不要相信它！很少有研究者的思想如同报告中写的那

① B. F. Skinner的“科学方法中的一个历史事件”（1959）是一篇关于研究中一些偶然过程的有趣的文章。

样有逻辑性。研究者所作出的很多决定都是基于直觉或内部直觉。他们的错误始于拙劣的猜想。他们会因错误的理由而选择了正确的实验方法，或因正确的理由而选择了错误的实验方法。

不幸的是，尽管在大多数情况下心理学实验实际上是一个令人激动的、无序的、偶发性的寻宝过程，但很多学生仍然对心理学实验感到厌烦，因为他们认为心理学实验既枯燥又单调。在你尝试做自己的第一个实验之前，你对心理学实验确实是知之甚少①。让实验报告简洁的最主要原因就是为了节约时间和空间。尽管去查阅你所有同事的错误是件很有趣的事，但你没有时间，期刊也没有空间满足你的奢望。实验报告是为了有效地传递信息而设计的，而不是为了取悦读者②。

省略某些东西

有一种整理实验报告的方法是剔除一些实验和分析③。假设有一天你运气不好，而在你设计一个系列实验中的第 3 个实验时，觉得自己缺少灵感，头脑也很混乱。此时没有人对你的生活条件、内脏器官或大脑状态感兴趣。因此，你对这个实验很失望。我并不想知道这些。你也不想写这些，所以，也不写这些。不写这些对于科学没有任何损失，对我也没有任何损失，而你却保留自己面子。然而，你必须确定，自己没有因为实验结果不支持你的假设而丢弃一个好的实验。这样做并不属于简化伎俩！

假如整个实验并不能说明什么问题，那么，忽略它们不仅是可以接受的，而且有些时候放弃一些数据分析细节的做法也是正确的。分析数据的方法可能有很多种，你使用了所有的分析方法。尽管你应该如实报告所有分析，但你仅需要将那些最有代表性和包含最重要信息的数据分析结果进行详细报告④。

重新组织

特别是在开展探索性研究时，你可能会发现，一个实验的结果显示它不应该是系列中的第一个实验。你可能会发现这个实验需要一些实验数据的支持，因此，需要做一些预备实验。在这种情况下，你不需要告诉读者“由于实验者的误判和缺乏预见力，下面的实验顺序出现了混乱”。你可以将它们按照最有逻辑的顺序进行报告，不管这个顺序是否与你实验时的顺序相一致。数据就是数据，你应该尽可能以最有效的方法报告，因为扭曲真相并不是对科学的追求。

重新表述

最后，通常情况下，重新表述实验中潜含的理论是可以接受的。有时候你出于某种理由做了一个实验，后来发现对这个实验还有更好的解释；或者你发现其他研究者所做的实验对你正在做的实验有不同的解释。这种情况下，你不得不决定如何

---

① 像爱一样，阅读是实践的贫乏替代物。

② 很多教科书作者也这样认为，但他们从未从中找到乐趣！

③ 或者把它们放在脚注，没有人会读脚注。

④ 请注意：我并不是赞同做一个多重检验，然后，挑选出那些产生显著性的结果。如果那样的话，你是在歪曲显著性标准。（见 12 章）

让你的结论与这些新信息整合一致,以便对现有知识体系有所贡献。不幸的是,你的理论与这些新信息完全不相符合,因此,你不得不重新设计规划。然而,你常常可以通过改变关注点或重新解释结果让自己的实验符合修正的理论。在报告研究结果时,你也没必要增加读者负担,让他们再去阅读那些陈旧的理论解释。请再一次注意,你在伦理上的主要考虑仍然是采用有效的方式将研究结果纳入已有的知识体系。

在本章中我们不可能讨论你在实验中面临的所有道德问题。在有些情况下,你会发现,很难决定某些特殊的行为是否算是公正地对待科学。当问题出现时,你可能会希望与同事讨论,他们可能会提出一些你没有想到的其他选择。在最后,由你自己作出决定。如果你遵守为科学知识体系添砖加瓦的道德准则,那么,你就不会使用肮脏的伎俩,并很少施展令人生疑的诡计。

小 结

因为科学是个不断增长的知识体系,任何妨碍这个知识体系高效增长的行为都是不道德的。我们在很多方面不可能做到完全诚实。我们可能会热衷于伪造结果、剽窃或篡改履历等肮脏的伎俩。我们也可能使用可疑伎俩,例如,在实验设计中没能有效地控制混淆变量,或在数据收集过程中错误评定反应和误读指导语。在数据分析过程中没有完全满足检验的前提假设和不恰当地剔除被试以及歪曲图表,将系列实验拆开,进行单个报告,所有这些都是不可接受的。出于效率的原因,实验结果的报告形式和整个实验过程并不完全对等是可以接受的。例如,我们可以省去那些对报告没有价值的实验和分析,或者我们可以重新编排实验顺序和重新表述理论,只要这些行为有助于提高实验报告的效率。

6

如何找到已经做过的研究

Polonius:你读到了什么,我的主人?

Hamlet:词,词,词。

威廉·莎士比亚

但是,我们为什么要忍受这些学究们,这些人不厌其烦地将智力游戏小题大做?

S. I. Hayakawa(1978)

社会科学研究的殿堂已被彻底毁坏。它散落在成百上千的杂志和学位论文中。即使它不能成为一门科学,但为了其中仅存的共性,也应该对这些碎石进行仔细查看、精挑细选。

G. V. Glass, B. McGaw, & M. L. Smith(1981)

或许在你读第3章——关于获得一个实验构想的时候,一个极棒的主意会在你的头脑中闪现。或者那个主意像小猫一样踏着轻柔的步伐向你走来。不管怎样,我希望某些有趣的实验设想已经开始形成了,你正在迫切的希望开始你的实验。然而,在你开始正式的计划之前,你应该意识到,你的伟大的实验构想可能早已经被别人实现了。

为什么要查文献

虽然心理学是一个相对年轻的科学,但每年发表的文章也超过50 000篇。虽然其他研究者可能还没有做你正在做的事情,但是,非常可能的情况是,在心理学短短的历史之外,一些人已经做了类似的事情。重复别人的实验有悖于生产性原则,除非你认为已发表的结果不可信。

你或许也发现,研究其他人如何解决相似问题是非常有帮助的。你也许不熟悉他们使用的实验技术。你可能还发现,其他研究者已经发现了大量缺陷,你当然不愿意浪费时间再去发现一次。



科学的目标是构建一个有组织的知识体系,而不是通过科学家做的一些细小孤立的实验,形成一个随意的事实集合。因此,确定其他人已经做过了哪些研究的最重要的原因是,让你的发现能够被纳入已存在的知识体系之中。当你已经完成了自己的研究之后,你必须说明的不仅是"它是如何来的",而且还应该包括"它适合在哪里"。为了了解你的研究工作适合的地方,显然,在你的研究开始之前,必须熟悉这个知识体系。本章就讨论如何通过文献搜索①找出在知识体系里都已经有了些什么内容。

当你在图书馆里查文献时,应该将你找到的东西作记录。每当你找到一篇文献或者一本可能有用的书时,都应该将重要的观点记录下来,并写下完整的参考信息,

① 科学家习惯称之为文献搜索过程,尽管在同事看来你可能并没有研究这些文献。

包括作者的名字①,作品的名称,杂志或者书的名字、日期、卷号、页码等;对一本书而言,还应该写下出版商的名字。如果你在自己的研究中引用这些文献,上述所有信息都非常重要。有人发现,用索引卡片可以帮助标记每个参考文献。自动搜索是另一种优先使用的方法。正如我在本章后面要讨论的那样,自动搜索很实用,它的一个非常有用的特点是你可以获得每一个参考文献的完整记录。你可以很容易地将它们打印出来。不论你如何做研究,第一次搜索文献时越有条理,那么,你在搜索文献方面浪费的时间就越少,即使你把参考文献的信息写在丢弃了很长时间的口香糖纸后面。如果你找到了几篇特别好的文章,你应该把它们全文拷贝并妥善保存。在开始写作时,你可以将它们作为写作模板。

虽然文献搜索并不是一个特别困难的工作,但是却非常耗时;而且也不会使人特别激动,因为这个过程你看到的文章实在太多。然而,掌握文献绝对是必须的,它是你的科学知识体系中的一部分!当你把你花费了宝贵时间做出来的结果向人们展示的时候,却听到有人评论道:"当然,你熟悉 Klip & Klap(2006)吧,他们去年也做了同样的工作!"没有比这更令人尴尬的了!

资源的时限性

如果你对心理学中可能用到的研究资源完全不熟悉,如书籍和文章,你可能就无法知道从哪里开始你的研究。首先,你需要对所能获得的资源有所了解,并且知道每个资源是如何更新的。为了做到这一点,让我们以一个在科学界发表的典型实验为例。图 6-1 用时间进度总结了这一过程,其中的 0 点代表研究者开始一个研究项目的时间②。收集数据后,研究者可能向自己的一部分同事报告研究结果。假设研究者没有被嘲笑,他可能决定去参加一个学术会议,如学术年会,并宣读该研究结论③。假设这些更挑剔的听众提供了一点支持,研究者可能就会决定将实验成稿,并将它投往一个学术杂志。如果文章被接受,它将在 9 个月至 1 年之内出现在杂志上。随着杂志出版,《心理学摘要》(*Psychological Abstracts*)将会发表这篇文章的摘要。如果这篇文章足够重要,它可能会出现在《年鉴》(*Annual Review*)上,被其他杂志引用,并且可能被《心理学通报》(*Psychological Bulletin*)这样的杂志收录。最终,在几年之后,教科书的作者可能会注意到你的研究并把它收录其中。

图 6-1 表明,科学交流过程存在滞后现象。如果你使用图书馆,你首先得到的是实验结果是杂志上发表的文章。正如你所见到的那样,从这一点上来讲,就已经失去了很多时间,因为某个研究可能至少 3 年以前就开始了。如果你现在开始研究并经历同样过程,那么,其他人将必须等待3年或更长时间才能在一本杂志上看到

---

① 刚进入心理学领域的人可能会发现,实验心理学家在谈论实验时总是说作者而不是被试。如果老师在上课时说到"Carothers, Finch and Finch (1972) 的眼光或观点与 Peterson, Bergman, and Brill (1971) 一致,"那么,他们说的不是律师事务所,而是实验者。

② 该图以较早的研究为素材。可能由于出版技术的发展,时限变得更加短。然而,我知道目前没有其他关于这方面的研究。我相信,程序的时间顺序直到今天都没有变化,因此,这个图仍然能够帮助我们了解论文发表的过程。

③ 这是教授们在没有课的时候要去的地方。你可能认为他们去度假了。

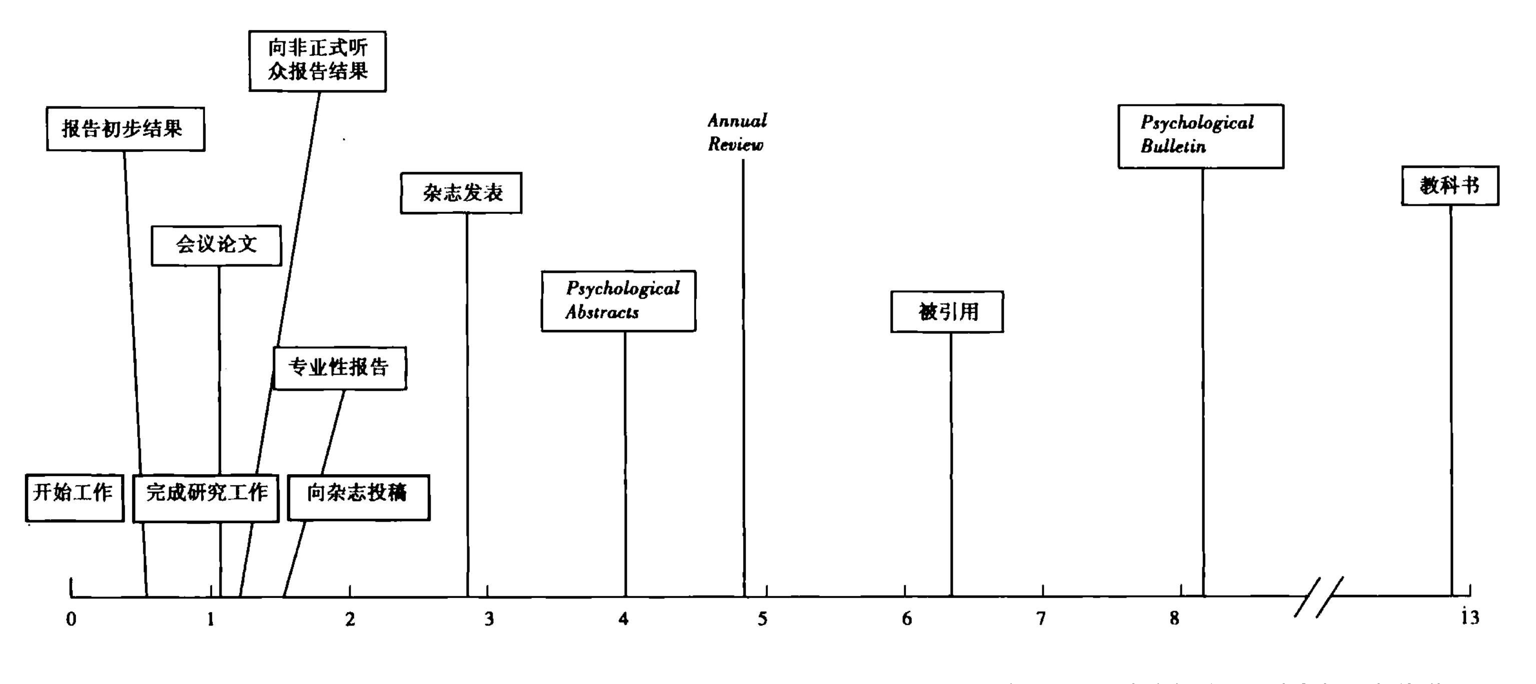


图 6-1　一个研究从产生到进入心理学文献的过程。粗线条框内说明了研究者从开始和完成研究到向杂志投稿。细线条框及竖线说明了从研究开始到它进入各种心理学文献所需的年数。来源：修改自"Scientific Communication: Its Role in the Conduct of Research and Creation of Knowledge," by W. D. Garvey and B. C. Griffith, 1971, *American Psychologist*, 26 353. Copyright 1971 by American Psychological Association.

你的研究结果(甚至美国邮政周转时间都比这个更好!)。由于需要避免这种延迟,不到七分之一的研究源于正规文献资源,如杂志上的文章(Garvey & Griffith,1971)。大多数研究构想是来源于特定领域科学家们进行的非正式交流,如会议上的报告、讨论列表和网页。然而,作为一个新手,你没有机会接触这些非正式交流的信息,因此,你不得不暂时求助于正式资源。如果你仍然继续在某一研究领域工作,你将会发现其他人在该领域的工作,并逐渐会与同行们建立私人交流;那么,你得到的信息将领先于杂志,成为"新的"研究者。

在下一节中,我们将更加细致地讨论正式资源,比较每一种资源的优劣,确定相关资源的定位。让我们从书籍开始,打开通往最新资源的路。

正规资源

书 籍

因为书籍只是包含几年前做的研究,所以,你可能会认为它是最差的文献资源。然而,这种巨大的时间延迟使得书籍却在某种程度上成了最好的资源。在研究完成和被某种资源发表之间存在一种重要的过程使得,即这项研究的重要性和质量经受住了时间的考验,因此,它才会出现在一本书中;它已经与其他研究一起被整合到一个较为完整的知识体系中。书籍的重要价值在于,作者把某项工作收入到一本书中,是因为他认为这项研究做得很好并且非常重要,它适应不断增长的知识体系。作者已经为你的研究做了很多事情,只是知识有点过时了。

最近出版的书籍是一个开始进行文献搜索的好地方,其中涉及了你感兴趣的一般性的研究问题。如果作者已经做了很好的工作,你就可以十分确信你有了一份很有用的总结,这份总结包括了自该书出版之前的 13 年来的最重要的研究。现在你的工作就简单了,你只需要找出来在这 13 年中该领域都开展了哪些研究。

这种方法的一个问题是作者对以往研究是有选择性的,它不可能涵盖该领域自心理学诞生以来的所有研究。每一个作者都有所侧重,或者偏重于某些理论,或者以方法为重点;他们很可能就是按照这种偏向选择相应的研究。此外,大多书籍并不像杂志那样接受同一水平的同行评议。同行评议的意思是若干个德高望重的研究者阅读了这些材料,并给出了他们的评价。大多数主要的杂志需要对所有文章进行评议并择优出版,但书籍的出版商很少这样做。因此,一定要确定你可以信任作者的学术成就和偏向,努力找出几本书籍的一致性;至少尝试发现作者的侧重点。

我想你应该知道如何在图书馆里找到书籍。如果你不知道,那么,就去问图书管理员如何查找,这比我描述更加有助于你进行文献搜索。图书管理员那里有全部现存书籍的电子目录。你可以通过输入作者的名字、书名或你感兴趣的主题进行搜索。如今,通过计算机网络可以搜索其他图书馆所收藏的书籍,虽然得到那些书可能并不容易。馆际互借在这里就派上了用场;但是,请一定不能等到你已经收集完了数据,在写结果的最后一分钟才去使用馆际互借。那时候再找到你想要的书就太晚了! 最后,你可以使用计算机检索数据库找到想要的书,这一点我会在本章后面谈到。这种情况下,你当然不能保证你想要的书在本地图书馆里都有收藏。

在你去图书馆之前,你可能也会查找与题目有关的介绍性的心理学书籍。大多数基础性的教科书会给出一个建议阅读的书目。你也可能会与所在的心理系中能够在你感兴趣的方面给予指导的老师谈谈该怎么做。他/她可能很愿意给你介绍基本参考书。最后,《美国心理学联合会图书馆使用:心理学手册》(Reed & Baxter, 2003)应该对你在学习心理学的图书馆使用技巧方面非常有用。

综述文章和书

其他一些资源也试图对心理学某一领域的研究进行总结与整合。这些资源比课本更有时效性,研究者可以在很短时间内掌握以往的一些研究。其中之一就是APA美国心理学会出版的一本杂志,叫《心理学通报》(*Psychological Bulletin*),这本杂志刊登"评价性和整合性,并对科学心理学中重大的方法论问题的解释的综述",下面是从一些杂志的不同期中摘抄的题目:

"唤醒和倒U曲线假设:对Neiss的'概念重新形成的唤起'的评价"
"婚姻的归因:回顾和评价"
"数学行为中的性别差异:元分析研究"
"酒精对人类攻击性的作用:交互作用研究综述"
"后见之明:对已知结果的事件的偏差性判断"
"哲学和心理学的因果观念"
"抑郁症的心理治疗:控制结果研究的综述"
"抑郁父母的孩子:综合评述"
"科学和道德:价值观在科学和科学研究中的道德现象的作用"

由此可见,这些文章的题目都比课本的题目要窄。一篇通报的文章可能以以前的总结性文章作为其起点,而不是把心理学的开始作为起点;因此,它缺乏整体性的调查。尽管如此,近期的综述文章一般来说比书更具有时间性,它通常包括了5~8年来的研究内容。

对于与你正在计划做的实验类似的实验而言,综述文章无法提供给足够的细节。这种情况下,通过参考文章后面引用的原始文献你可以迅速找到那些重要的参考文献,并且帮助你确定你的实验与过去的研究是否相互呼应。

另一个研究综述的资源是《心理学年鉴》(*Annual Review of Psychology*),它由年鉴公司(Annual Review, Inc)出版。这本书中的文章每年都不一样,这由编辑部决定。每一章都由一个作者撰写,该作者被公认为是该领域的专家,他的工作就是总结与整理以往有关该题目的相关研究。题目通常会比《心理学通报》(*Psychological Bulletin*)更加宽泛:

"人格"
"发展心理学"
"空间视觉"

而有些题目就很窄①:

① 你注意到了这种有趣的关系了吗?话题越广,则题目就越短。

"干预技术:分小组"
"社会和文化影响精神病理学"

近年来,出版了很多编写的书。其中有些书总结了心理学某个领域的最新研究工作。每一章通常由一个研究者撰写,它对一些更新的甚至更窄研究领域的研究进行综述。这些章节与综述文章相似,如果你发现某一章与你的研究相关,你就可以节省很多研究时间。这种书的出版周期比课本短。实际上,很多都是被"桌面出版"制作出来的,其中的排字、编辑和生成终稿的时间都大大缩短。这种情况下,一些研究可能只有 1 ~2 年就出版了。然而,必须再次重申,这些书并不像杂志那样通过了同行评议。

期刊文章

心理学期刊形成了我们科学的支柱。它们被称为基本资源,因为它们展示的基本结果是研究者自己获得的,而不是由综述作者这样的第三方获得的。为了彻底进行文献搜索,你必须使用杂志期刊。它们是最新的正式资源,是几年前刚刚完成的真正的研究。这样,虽然文章作者试图将它们的工作整合到现存的知识体系中,但他们的努力也只是部分的成功。因为他们经常不知道当时正在做的研究与其他研究都有什么联系。因此,你不得不自己设法整理这些工作,以使得你的研究形成一个有序的知识体系。我不可能在这里列出所有心理学相关的杂志。很多专业的组织为他们的会员出版期刊,某些出版公司也赞助出版独立期刊。为了给你一个关于现行期刊的印象,以下是一些杂志名称:

《美国心理学杂志》(*American Journal of Psychology*)
《动物学习与行为》(*Animal Learning & Behavior*)
《听觉学》(*Audiology*)
《行为和脑的科学》(*Behavioral and Brain Science*)
《行为神经科学》(*Behavioral Neuroscience*)
《认知》(*Cognition*)
《认知心理学》(*Cognitive Psychology*)
《现代心理科学方向》(*Current Directions in Psychological Science*)
《发展心理学》(*Developmental Psychology*)
《变态心理学杂志》(*Journal of Abnormal Psychology*)
《应用心理学杂志》(*Journal of Applied Psychology*)
《认知神经科学杂志》(*Journal of Cognitive Neuroscience*)
《比较心理学杂志》(*Journal of Comparative Psychology*)
《行为实验分析杂志》(*Journal of the Experimental Analysis of Behavior*)
《实验心理学杂志:动物行为过程分册》(*Journal of Experimental Psychology: Animal Behavior Processes*)
《实验心理学杂志:普通分册》(*Journal of Experimental Psychology: General*)
《实验心理学杂志:人类知觉与行为分册》(*Journal of Experimental Psychology: Human Perception and Performance*)

《实验心理学杂志：学习，记忆与认知》(*Journal of Experimental Psychology: Learning, Memory, and Cognition*)

《记忆与语言杂志》(*Journal of Memory and Language*)

《人格和社会心理学杂志》(*Journal of Personality and Social Psychology*)

《学习和动机》(*Learning and Motivation*)

《记忆和认知》(*Learning & Cognition*)

《动机与情绪》(*Motivation and Emotion*)

《认知与心理物理》(*Perception & Psychophysics*)

《心理学记录》(*Psychological Record*)

《心理学综述》(*Psychological Review*)

《心理学科学》(*Psychological Science*)

《实验心理学季刊：A 刊，实验心理学》(*Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*)

《实验心理学季刊：B 刊，比较和生理心理学》(*Quarterly Journal of Experimental Psychology. B, Cpomparative and Physiological Psychology*)

会 刊

在心理学的一些领域，来自科学会议报告的会刊也非常重要。例如，人的因素和工效学协会出版会刊，既有电子版也有纸质版的，就包括大部分在年会上发布的文章。而这些会刊文章一般比杂志文章简短，它们由一组专家进行评议，并且被认为是该领域最重要的信息来源之一。因为篇幅的限制和缺乏足够的同行评议，会刊文章通常都被认为比杂志文章价值要低些。但它们确实为研究结果及时进入科学文献系统提供了一种渠道。

技术报告

作为一种心理学文献资源，技术报告经常被忽视。但它们在某些领域却是非常有用的。当联邦政府支持研究工作，尤其是国防部的研究时，研究者通常被要求以技术报告的形式进行汇报。这种报告与杂志期刊相似，但通常要求对程序和装置部分进行详尽描述，有时甚至要列出数据。支持项目的政府机构会自动地向该机构内从事相似研究的其他研究者发布技术报告。

大约有十分之一的研究者撰写技术报告，但只有大约三分之一的报告会在杂志上发表(Garvey & Griffith, 1971)。大多数图书管理员通常并不订阅技术报告，因为他们会迅速填满书架，并且很难识别和系统地进行组织分类。获得军方研究基金或合同的研究者会得到登载所有技术报告摘要的月刊。《心理学摘要》(*Psychological Abstracts*)也列出了很多这样的报告。不幸的是，技术报告经常很难获得。若要购买它们，你必须发送一个请求到位于弗吉尼亚的 Alexandria 的国防文件资料中心(Defense Documentation Center)，并且你必须知道文件的编号和你想要的报告的标价。

查阅技术报告对某些领域来说纯属浪费时间。然而，如果你正在一个被主要的政府机构支持的研究领域工作，技术报告就是一项非常有价值的信息来源。政府支持的一些研究案例包括汽车驾驶员安全、个人训练和选拔、复杂机械的操作控制以

及人类决策过程。

电子出版物

即使有人会觉得我过于信任电子出版物，我仍然将它看成是正式资源。电子出版物的范围非常广，从研究者最近在网页上发布的手稿，到经过同行评议而发表在电子杂志上的文章，这些文章的质量等同于印刷的杂志文章。就这一点而言，有时候很难确定一个人找到的文献的质量。有些专家预测，电子出版物将会成为主导的科学交流形式，最终完全消灭纸质出版物。但是，至今尚未实现这一目标。电子出版物的优势非常明显。人们能够立刻从互联网上获得文章，并且只需花打印费就可以获得纸质拷贝。研究者能够邮寄文章的手稿，并且根据读者的意见对它们进行修改。

然而，电子出版物也存在问题。谁来保证这些文章的质量呢？出版杂志的专业协会仔细挑选编辑，这些编辑将这些稿件发给权威的同行评阅者。有些情况下，只有10%到20%的投稿会被选中出版，而且必须经过一位文字编辑修改和仔细的校正后，文章才会被出版。而在开放的网页上，任何人都可以制作一个网页并发表一篇研究报告。当不同的稿件被发到网页上时，第二个问题出现了，即哪一个版本是最终版本？出版物（尤其是杂志文章）形成了科学的基础构件，作为科学家，我们需要有某种方式判断一篇文章是否可以作为独立的构件添加到科学体系之中。最后，就是钱的问题。专业学会通常是科学的保有者。他们提供了基础结构允许科学事业有序扩展。这种基本架构很昂贵，经费很大部分来自图书馆和科学家个人的订阅费用。如果文章在网页上免费出版，那么，从哪里得到经费以支撑科学的基本架构呢？

这只是一些已经知道的问题。所以，美国心理学会——一个拥有最大数量、享有盛誉的期刊的出版商，发布了一个政策，即将不会出版任何已经在网络上贴出的稿件。它认为，如果一篇稿件已经被贴了出去，那就意味着该稿件已经发表了。这并不意味着你不能把稿件通过电子邮件发给人们审阅。然而，如果你希望最终在标准的科学杂志上发表自己的研究，那么，你就应该遵循这项政策。



文献搜索

你还应该在做文献搜索时就想到这些问题。你可能会在网上发现一些重要的文章。特别是那些贴在声名显著的研究者的网页上的文章更可能被人们参考并在自己的文章中引用。然而，互联网上也不乏垃圾。你应该带着批判眼光去审视你用非科学的电子搜索引擎发现的信息。网络上没有强制的质量控制，所以，你不得不自己做质量监控。在你引用网上的信息之前，请准确无误地核查作者的信誉，并咨询本领域有经验专家的意见。

查找正式文献

查遍海量杂志和书籍是不可能的，这一点可能使你大伤脑筋。我不会责备你；你可以耗尽毕生精力在图书馆里浏览各种出版物，但你仍然会跟不上出版的速度，并被甩得越来越远。值得庆幸的是，美国心理学会救了你。他们的服务之一就是 *PsycINFO*，他们致力于制造产品来帮助研究者寻找他们需要的文献。自从 1927 年以来，美国心理学会出版了《心理学摘要》，它包含参考文献和文章的摘要。直到最近几年，学生在搜索文献时还要去图书馆里寻找不同卷本；然后，通过关键词或者作者名字，以手动方式找到与他们主题相关的文章。这个工作很沉闷，但非常必要。

在当今，电子交流方式的普及大大简化了这项工作。除了文摘之外，美国心理学会还出版了 *PsycINFO* 电子数据库。很多大型的图书馆都购买了相应的租赁协议，使得无论在图书馆的终端上，还是在远程都可以登录使用。例如，在我们大学中，学生和教员可以通过个人计算机在房间、办公室等场所进行文献搜索。这个数据库包含超过 1 500 万个心理学文献摘要，时间跨度从 1887 年至今，而且每月还会增加 5 500 个新的参考文献。它们包括文摘、学位论文、报告、英文版书籍的章节以及其他学术文献。

图 6-2 是一个数据库条目的例子。三个基本部分分别是书目引用、总结和标准主题索引。这个书目引用包括题目、作者和来源。如果你想在文章中引用这篇文章，那么，这些是参考文献部分必须引用的信息。总结通常是一篇摘要，对杂志文章来说，通常由作者撰写。总结并不对研究进行评价，仅仅是描述而已。标准的主题索引使用关键词和描述词。请注意，你从条目得到的所有内容只是对一个研究的简要描述。它为你提供了是否对这个研究感兴趣的足够的信息，但这并不能替代对参考文献原文的阅读。为了阅读整个文献，你必须从图书馆得到文章、书籍或章节。很多大型图书馆都收藏了标准的杂志。如果图书馆没有你要找的文章，图书管理员将能告诉你如何通过馆际互借获得。现在也有文献全文传递资源，它可以通过互联网向你发送杂志文章，这通常需要付一定费用。有些图书管理员也订了 *PsycARTICLES*。你可以通过这种服务从 APA 杂志获得全文，目前，它可以提供能够追溯到 1894 年的 24 种 APA 杂志上的全文。

你如何找到与自己研究题目相关的条目呢？这里有几个方式可供参考。如果你找到了某个感兴趣的题目，你可以用关键短语或者描述词进行搜索。我就是用这种方式发现的图 6-2 的条目。我正在计划做一个实验，探讨可得性启发（availability heuristic）可能影响人们对原子能发电的支持率。这种可得性启发是指我们的观点会受到与观念相关的物质的影响。例如，我可能持有这样的观点，即人们死于鲨鱼的次数高于被黄蜂蛰的次数，因为我更容易想起鲨鱼袭击。而实际上，黄蜂蛰死人

ACCESSION NUMBER: 1997-05223-007
DOCUMENT TYPE: Journal-Article
TITLE: The availability heuristic: Effects of fame and gender on the estimated frequency of male and female names.
AUTHOR: McKelvie,-Stuart-J.
SOURCE: Journal-of-Social-Psychology. 1997 Feb; Vol 137(1): 63-78
ISSN: 0022-4545
PUBLICATION YEAR: 1997
ABSTRACT: In 2 experiments, Canadian undergraduates heard a list of 13 male names and 13 female names; then they estimated how many male and female names there seemed to be. In Exp 1, the list consisted of 26 famous names or 26 nonfamous names. Both male and female participants gave similar estimates for the number of male and female names, contradicting hypotheses of a bias toward males or toward one's own gender. In Exp 2, where the list contained names of famous men and nonfamous women or names of famous women and nonfamous men, participants gave higher estimates for the gender that was famous. This result confirmed A. Tversky and D. Kahneman's (1973) fame availability effect and showed it to be moderate to large in size. ((c) 1997 APA/PsycINFO, all rights reserved)
KEY PHRASE: fame and sex of name, estimated frequency of male and female names, college students, Canada, test of fame availability heuristic
MAJOR DESCRIPTORS: *Estimation-; *Fame-; *Human-Sex-Differences; *Names-
MINOR DESCRIPTORS: Adulthood-

图 6-2 选自 *PsycINFO* 数据库条目的典型数据样例

的比率更高。我们在实验中让被试给出原子能发电的优势或劣势;然后,让他们对核电进行评价。我们预期,通过回忆优势或劣势会使得优势或劣势更加容易想起来,因此,会使支持率发生改变。在这种情况下,我的兴趣在于找到和可得性启发相关的参考文献,这样我就可以输入这些词进行搜索了。

我得到一条信息,该信息告诉我搜索到了 50 篇文献。于是我开始浏览这些条目,如图 6-2 所示。我读了这个条目中的题目和摘要,觉得与我的兴趣还不够紧密,不值得去阅读整个文章。为什么会搜索到这个文献? 原因在于,在该条目中,题目和关键词都包含了“可得性启发”一词。这个名词可以出现在条目的任何地方,因此,这个条目就被选中。我其实可以通过确定特殊搜索域来限定搜索。例如,我可以只搜索关键词域。在图 6-2 中只显示出了 11 个搜索域,但实际上有 89 个可能的搜索域①。通过这种方式,你能把你的搜索限定在杂志文章范围内,或英文文献,或成年被试等。用这种方法,你不至于搜索太多条目。

搜索的具体过程过于琐碎,就不在这里多说了,而且搜索引擎也频繁变化。你可以从管理员那里学习这一过程,也可以从一些小册子上,甚至从 APA 的网站上(http://www.apa/psycinfo/)。但是,我还是要介绍一下通常的搜索步骤。首先,应该缩小你的研究问题。例如,假设你想知道,“人们对使用计算机焦虑吗?”你应该明确问题中的独立概念。例如,焦虑和计算机。然后,你应该使用在 *PsycINFO* 上面提供的电子分类汇编找到合适的描述词。这个汇编会让你扩大或者缩窄所需要的专有名词范围,从而寻找符合你目的的名词。对于每一个名词,你都应该做这一步。例如,焦虑也许应该包括恐惧或惊恐。你接着就需要将你的描述词使用“与(AND)”“或(OR)”或“非(NOT)”连接起来。此时,你必须非常小心,因为这些词的含义相差很大。“与”表示所有你想要搜索的词必须包含某一条目中,如既包含

---

① 你在 http://www.apa.org/psycinfo/网页上可以看到更多的搜索域。该网页也会详细说明如何进行电子搜索。

"焦虑",也包含"计算机"。"或"是一种宽泛的多重搜索,它是指所要搜索的条目可以包含两个描述词中的一个。"非"用在你想要排除某个描述词的搜索中!

一旦搜索完成,屏幕上会告诉你找到了多少个条目。如果只是很少或没有,不要马上就觉得本领域内这方面的研究没有或很少。你应该对你问题相关的概念再次进行评价与分析;或者查看这些仅有的结果,以生成新的描述词。如果搜索到了几个有用的参考文献,要看一下条目中的描述词,检查这些词对你的研究是否合适;如果合适,就将它们加入你的研究中。尝试将你的描述词用不同方式连接起来,看一看是否影响找到的条目。换句话说,如果你已经搜索到了上百条条目,快速浏览一下这些条目,检查是否有你根本不感兴趣的研究领域,然后,可以通过使用"非"或者其他方式的连接描述词,以滤掉这些条目。

查阅电子数据库的另一种方式是使用作者而不是描述词。你也许知道一个或者几个作者经常在你感兴趣的领域内发表文章,或者你已经在查文献的时候发现了这样的作者。你可以把这些作者的名字输入到数据库的作者域中,看一看这些作者还发表了哪些文章。例如,我在检索可得性启发的文献时,我知道这个名词是由 Daniel Kahneman 和 Amos Tversky 提出的,他们已经发表了其他一些有关这个主题的文章、书籍的章节和著作。因此,使用他们的名字在数据库里进行检索,就可以发现一些更为相关的参考文献,当然也会有很多不相干的其他文献。一定要使用名字的所有变式,即包含和不包含名字的首字母,因为很多情况下名字是用不同的方式列出来的。

在你从各种各样的搜索中终于得到了一个合理的条目列表后,需要对它进行排序,挑出最合适的文献;然后,把它们标记出来,以便后来打印或下载整个条目,或者通过电子邮件发给你自己。这个列表将提供了你需要的信息,并且可以通过它们找到原始文献、书籍或你想要读的章节全文。以后,你可以从列表中选择书目的条目,从而生成研究报告的参考文献部分。

我已经非常详细地介绍了如何使用 *PsycINFO*,因为它是心理学界最广泛使用的数据库。然而,心理学家发现,还有其他一些科学文献数据库也很有帮助。例如,特别是对偏向临床/医学方面的心理学科来说,*MEDLINE* 就是非常好的数据库。使用 *MEDLINE* 与使用 *PsycINFO* 很相似。

假设你对躁郁症(bipolar disorder)感兴趣,这种疾病是心理学家更为常见的心理障碍。你可以在搜索框内输入 bipolar。然后,你对搜索进行限定,如在题目、作者名字、杂志名称或任何需要的地方填上适当的词。还有搜索的时间跨度、结果如何排列,以及是否需要只有引用、摘要或是全纪录等。

我输入 bipolar,并限定在 1995 到 2002 年发表的文章,搜索结果列出了 250 个条目。为了使搜索更加高效,你可以用增加关键词的方法缩小搜索范围。假设你真正感兴趣的是被诊断为躁郁症的患者的自杀行为,你应该在关键词中输入 bipolar 和 suicide。当我这样改了之后,查到的条目从 250 降到了 15 个,这就容易多了。

我提倡你去图书馆通过询问或电脑查找那里有哪些数据库。你会发现,除 *PsycINFO* 之外,其他数据库会拓展你的搜索范围,特别是在你的研究领域与心理学外的某个领域交叉时更有效。

用"树图"回溯搜索文献

还有一种与 *PsycINFO* 不同的文献搜索的方法,它不如 *PsycINFO* 那样好用。然而,这种方法可以确定你以前的搜索中是否遗漏了某些关键文献。我把这种技术称为树图回溯搜索参考文献。第一步是找到与你感兴趣题目最接近的文章;这篇文章就构成研究树的主干。在这篇文章后面找到它的参考文献。参考文献中的大多数应该也与你的研究兴趣相关。(要是幸运的话,你所需要的大部分参考文献已经包含在列表之中。)这些文章也各自有一个参考文献列,你还可以采用同样的方式进行选择。随着参考文献的增加,你或许可以发现所有重要的文献,这些文献形成了树的次级分支。这种方法非常有用,但不要将它作为唯一的技术,并不是每一个作者都能够发现重要的参考文献。



树图文献搜索

"树图"前溯搜索文献

为了使文献搜索比较彻底,你可以像回溯文献一样进行前溯文献搜索。例如,如果你找到一个很重要的文献,它是几年发表的,你想了解它被哪些文献引用过;那么,你可以使用社会科学文献检索(Social Sciences Citation Index,SSCI),它是一个数据库,由科学信息研究所研发,每季度更新且每年度汇总。有近1 400种杂志,几乎覆盖社会科学的所有门类。现在 SSCI 既有电子版,也有纸质版。在任何一次搜索中,你可使用一篇关键文章,你可能想到所有引用了这篇关键性文章的文章;那么,为了搜索这些文章,你需要查找从该篇关键性文献发表以来所有引用过它的文章所在杂志的每一卷。在每一卷里面,除了这篇关键性文献外,还有一些引用过它



如何循环搜索

的其他文章。如果你做的是电子搜索,你可以立刻搜索到所有年代,像在其他数据库里一样,你可以标记上你感兴趣的条目,然后,将它们打印出来。目前,电子数据库也允许你按照文章的主题或来源进行搜索。

你也可以查找所有引用了最初文章的文献,然后,再回溯每一篇文献后面的参考文献,并不断循环;你也可以将一些最近得到的参考文献作为关键文献,然后,再前溯搜索。你可以不断重复这样的过程,直到你觉得自己已经掌握了所有的重要参考文献。

索要抽印本

科学家们在职业上的相互尊重,可以使你免费得到一些杂志的文献。当作者的文章出版后,他们通常会向出版机构订购 100 份左右的抽印本。只要还有抽印本,并且你非常礼貌地向作者索要,他们都会送给你一份。通常采用的索要方式是向作者发送一份明信片说,“非常希望你能送我一份你在______杂志上发表的题为________的文章抽印本,非常感谢!”你也可以给作者发电子邮件而不是明信片。如果知道作者在哪里工作,你就能发现他的电子邮件地址。例如,如果作者是一个全职教师,就直接去搜索引擎上寻找他所在的大学,找到该教师的大学网页。然后,浏览心理系的网页或者大学的地址名录,你通常会发现可用的电子邮件。如果你对作者的研究领域感兴趣,那么,当你与他们联系时,你可以顺便索要一些其他相关的文章。你要确保电子邮件里包括你的地址。作为礼貌,作者通常会送你一份抽印本。有些情况下,作者可能会通过附件给你发送文章的电子版。不要害怕发送这些抽印本请求。很多年轻的研究者正在努力熟悉某一特定领域,但没有经费资源购买自己的杂志,于是他们就发送抽印本请求。大多数作者会认为,这种请求与其说是找麻烦,不如说是一种恭维。

现行研究

Smithsonian 科学信息交流会(Smithsonian Science Information Exchange)提供了一个了解目前研究进展的途径。他们的文件中涵盖了所有行为科学领域的 14 000 余份研究项目的记录。所有这些项目都获得了像国家科学基金这样的机构支持。每一条记录都包括一个对当前工作的 200 字左右的描述。你可以订购一份包括一般题目的记录,例如:“失眠症”或者“酒精成瘾的行为治疗”。这项服务需要根据订购记录的数量付费。这个系统的缺点在于,它的费用高,并且只包含了被基金资助的研究记录。然而,这却是一个找到当前研究进展的很少的方式之一。

非正式资源

专业会议

如前所述,为了完全地跟上某个领域研究步伐,你必须熟悉交流的非正式资源。在杂志出版前的大约 15 到 18 个月,很多研究者都会利用在专业会议上的交流机会发布自己的研究结果。实际上,发表在主流心理学杂志上的大约五分之一的文章,

都曾经在 APA 大会上报告过(Carvey & Griffith, 1971)。美国心理学会每年都会组织全国性会议和 6 个区域性会议。此外,很多其他非-APA 的专业性组织,如心理学研究会(Psychonomic Society)、心理测量学会(Psychometric Society)以及心理科学协会(Association for Psychological Science)都会组织各种学术会议。

当然,你不可能参加所在领域的所有会议。如前所述,在某些会议之后,会议中报告过的文章会被编辑出版在会刊上,在大多数图书馆中都能找到这种会刊。另外,在大会之前,这些组织的会员会收到大会日程。你可能会找到所在心理系协会会员,从他们那里得到一份大会日程。一旦你发现自己感兴趣的文章,就可以直接向作者索要一份抽印本。阅读文章总要比听报告更能帮助你理解文章。

除了参加额外的娱乐活动外①,参加会议的真正原因就在于与该领域的其他研究者交流。虽然他们也有防备之心,但你也许能够发现他们将来的研究趋向。通过这种方式,你就能够填上图 6-1 中的"开始工作"和"会议文章"之间的信息鸿沟。

顺便说一下,如果你在某些讨论中获得了想要在文章中引用的一些内容,一定要将它记录下来,并标明时间;在获得这个作者同意后再使用。你可以将它作为"个人交流"进行引用。

研究小组

一旦你了解了谁在你感兴趣的领域里做研究,你会发现他们已经建立了一种非正式的交流途径,彼此可以互通信息。有些情况下,这是一群人,他们互送文章的抽印本,或者已经准备好的稿件,甚至是他们正在写的初稿。当今的互联网为这些小组成员之间提供了一种交往的方式。有时,小组成员可以通过电子邮件或者聊天室进行讨论。有的小组也非常欢迎其他任何感兴趣的人加入讨论。有些小组则较为严格,你必须得到邀请才能加入。一旦你对某个领域产生了兴趣,请你一定关注这样的小组的活动。它提供了一种非常有价值的方式,使得你能够及时获得最新的研究信息。

同事成员

千万不要忽略身边能够提供非正式帮助的资源——心理系的同事们。学生认为,教授太忙了,不会帮助他们,因此,不愿意接近教授;他们甚至认为,请求老师的帮助就是一种欺骗行为。恰恰相反,大多数人不仅十分愿意提供帮助,甚至把被询问当成一种恭维。这种帮助更像是教学的一部分,就像站在教室里面讲课一样。真正的研究者会去使用他们所能找到的任何资源,以帮助自己开展研究。科学需要团队的共同努力,其目标是构建知识体系,而不是研究者之间的对抗,或者学生与老师之间的斗争。因此,应该去尝试一下! 你会非常高兴而惊奇地发现,你的教授不仅愿意提供帮助,而且他们也非常博学!

尽管科学研究的主要文献来自正式资源,非正式资源在科学研究中也起着重要作用。它们提供了一个论坛,在这个论坛中,人们可以互相说对方愚蠢,但却创造出了许多新生事物。你的非正式的同事们可能会暗自窃笑,并告诉你错在哪里。而你

---

① 娱乐而已。

的正式的同事则不得不高谈阔论，向世人宣布你的错误所在。如果仅有正式资源，没有人能够有勇气去推动科学快速发展，我们只能变得保守而步履维艰。非正式交流提供的鼓励和友善的批评对我们将思想纳入正式文献尤其重要。

我已经尽可能完整地对文献搜索进行了讨论。我希望在讨论过程中并没有将它复杂化。许多新的研究者认为，文献搜索需要某种神秘的力量以及多年的经验积累。然而，如果你按照本章所描述的步骤，你就会发现，文献搜索是一种简单的令人愉悦的经验。

小　结

文献搜索对于确定你的实验构想是否被别人研究过、是否有人做过了类似的实验以及判断你的实验是否适合当前的知识体系等都是非常必要的。为了进行有效的文献搜索，你应该了解科学界的交流方式，以及各种信息来源的时间滞后特点。通常情况下，开始你的研究的最有效方式是从书籍中搜索你感兴趣领域的相关问题。书籍描述的内容往往是从心理学开始到距当前研究约 13 年之间的研究进展。综述文章则是当前研究之前 5 到 8 年的研究结果。杂志文章是文献搜索的主体。会议的会刊和技术报告是非常重要的信息来源，尤其是在应用领域。你可以在电子搜索引擎 PsycINFO 中，通过描述词或作者查找相关的文章、书籍和书的章节。你可以通过树状回溯反复检查你从最近杂志文章的参考文献中搜索到的文献。你也可以通过社会科学索引（*Social Science Citation Index*）进行树状前溯文献，以确定哪些文章引用了某个早期的文章。专业会议上宣读的文章、个人交流、抽印本，甚至同事之间的交流等非正式的资源都是了解当前和未来研究的非常有价值的方式。

7

如何决定需要操纵和测量的变量

我们认为，如果一个概念无法进行操作定义，那么，这个概念是没有意义的。

W. R. GARNER, H. W. HAKE & C. W. ERIKSEN (1956)

我们在第1章中了解到各种类型的研究;在第2章中讨论了实验的一般模型;在第3章中学会了如何获得实验的构想;在第4章、第5章中讨论了伦理问题。在第6章中,你可能已经学会的文献搜索方法比你想知道的还多。现在应该是坐下来工作的时候了,做实验心理学家应该做的事情——做实验!

在本章中,我们将讨论两个决定在计划做心理学实验的时候我们必须面对它们,它们最简单,也最为复杂,即选择自变量和因变量。

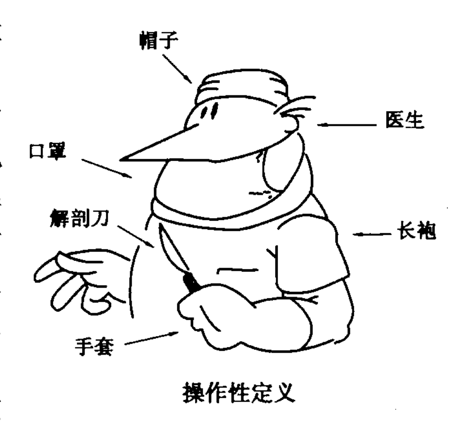
选择自变量

请回想一下第2章,其中曾经提到自变量是由实验者控制的。因为所有实验的目的都是为了寻找自变量对行为的效应,变量的选择是研究中你必须作的最重要决策。乍看起来,这个决策非常简单,对有些实验来说确实如此。例如,如果你想知道当声音作为报警信号时,人们的反应是否会更快一点。显然,自变量就是有无声音出现。然而,如果你想探讨孩子们看了暴力或非暴力的电视节目之后,他们的攻击性是否会变得更强,那么,自变量(暴力)可能就较难于进行定义。电视上的什么节目是暴力呢? 是星期一晚上的橄榄球比赛吗? 或者是跑路布谷卡通? 或者是说唱乐视频呢? 并不是每个人都会认同有关电视暴力节目的某一特别定义。

定义自变量

这里有这样一个问题,即一般公众和实验心理学家对于一个术语的理解存在着细微差异。实验心理学家应该对自变量和因变量进行操作性定义,这意味着他们必须对那些操作进行详细说明,其他人也能够像实验心理学家一样据此设定自变量。所以,一个操作性的定义有点类似于食谱,不同之处在于,用它的程序和成分制作的是一个变量,而不是蛋糕。

在第2章讨论的电视暴力实验中,操作性定义应该详细说明确定某个表演是否含暴力的步骤。例如,在你确定一个电视节目是否为暴力节目之前,你可以通过随机选择100人观看电视节目,如果75%的人认为这个节目是暴力的,那么,该节目就是暴力节目。另一种方式是设计一个调查表,项目包括:“有伤害他人的身体接触吗?”“有非法的行为发生吗?”以及“一个人的行为会让另一个人感到自卑吗?”等。你可以设定,每一个节目的调查表中应该至少有20%回答“是的”,才能考虑把该节目定为暴力节目。而且,这样的程序应该明确说明,其他的实验者必须采用哪些操作才能够满足你对暴力电视节目的操作性定义。



操作性定义



在操作性定义上心理学研究者比物理学家更难达成一致①。伽利略在决定两个不同质量的物体在真空中以同样的速度落下之前并不需要仔细考虑对质量的定义。而大量重要的心理学问题则需要复杂的操作性定义。例如，其母亲感情丰富的人们在婚姻中会更为成功吗？学生们从受欢迎的教授那里学的东西更多吗？工人的道德观影响工作效率吗？焦虑导致抑郁吗？在通过实验回答这些问题之前，你需要对感情丰富、成功、学会、受欢迎、道德、工作效率、焦虑和抑郁等词进行操作性定义。请试着对这些词进行操作性定义，你很快会发现心理学研究者面临的挑战。

对大多数需要进行操作性定义的概念而言，你会从文献搜索中发现，其他的研究者已经开展了工作。好消息是，如果他们已经做得很好了，也就意味着你可以节省大量工作。而坏消息是，如果你不同意他们的定义，你可能很难作出一个能够被人们接受的新定义。科学是相当稳健的，不喜欢快速变化。你可以想象，如果每一个研究者对每一个重要的概念都坚持使用不同的操作性定义，那么，情况会是怎样的混乱。科学的知识体系像巴比伦塔（the Tower of Babel）一样，每一个人都会说着不同的语言。所以，一旦某个概念已经被操作性定义，这个定义就会相对稳定；再去说服别人使用一个新定义通常是很困难的。所以，在你试图对实验中的专业术语进行操作性定义时，一定要去查一查文献，找出别人是如何对你想研究的概念进行定义的。

选择自变量的范围

一旦你已经定义了一个自变量，你还必须选择变量的范围。范围是指你所选变量的最高和最低水平之间的差异。例如，假设我们试图通过100个被试判断每一个节目是否具有暴力，从而定义暴力电视节目。我们可以选择两个暴力水平，即

---

① 操作定义一词由物理学家首先使用。在物理科学中，操作性定义得到了广泛认同，因此，物理学家在操作性定义方面比行为科学家所花的时间要短得多。

100%的人们都认为具有暴力以及没有人认为具有暴力。自变量的这两个水平就是最大可能的范围。

另一方面,我们可以按照超过50%的人认为具有暴力的节目就是暴力节目,而少于50%的人认为是暴力的节目就是非暴力节目。很显然,这种情况下的范围要小得多。

那么,我们如何确定范围的大小呢?不幸的是,我无法给你一个严格的固定规则,因为科学就像是一门艺术。然而,下面的一些讨论或许对你有所帮助。

要面对现实

首先,你选择的一个范围应该具有现实性,应该与研究的情境相适应。你应该避免"长柄大锤"效应,即把自变量的水平设定得过于极端,以至于行为很容易表现出差异。早期对大麻的医学研究就被长柄大锤效应所困扰。有些情况下,实验者给老鼠的大麻剂量相当于每天给人一卡车大麻的剂量!实验结果非常明确,但与实际情况相去甚远。

选择存在效应的范围

由于现实情况的限制,在你选择一个范围时,这个范围应该足够大,使得自变量对因变量存在的效能够表现出来。例如,如果你对室温影响分类任务中的手工操作灵巧性感兴趣,你选择的室温范围为23 ℃到25 ℃[①],你或许会得出一个虚假的结论,即室温对手工操作灵活性没有影响。

真实世界[②]的实验情境要求选择足够大的范围,因为实验者并不总是能够对自变量的水平进行完全控制。你可以大致选择一个水平,但是实际的水平可能会随试验而变化。例如,在第2章讨论的教学节奏的实验中,我试图改变说话的速度,包括低速、中速、高速。我选择的水平大约是每分钟100、125和150个字节。但是,由于我不是机器,讲话频率不可能非常固定,我讲话的频率肯定在某个水平附近不断变化。为了确定我实际的频率,我在讲课时进行了录像,并计算每秒的音节数目。幸运的是,慢节奏时的最快的讲课速度,也要比中等节奏中的最慢的讲课速度慢。因此,水平之间没有出现重叠。然而,如果我选择的范围较小,那么,在各种自变量的水平之间的差异就无法得到保证。因此,在有些非实验室实验中,你必须保证自变量有足够大的范围,从而使得自变量不同水平之间的差异不会被无法控制的变量掩盖。

做一个预实验

在某种程度上,确定自变量最适合范围的方法是猜测。某些情况下,当你查找文献时,你会发现,有些实验使用的自变量与你设计相同,这就给你选择合适的范围提供了参考。然而,如果你的实验是原创性的,以前没有人使用过与你相似的自变

---

① 华氏温度分别是73 ℉和77 ℉。

② 这里的真实世界是指用于研究实际问题的非实验室实验,并不是指多数大学中的人不真实。请生活在象牙塔中的人不要介意。

量，那么，你就应该做一个预实验（pilot experiment）①。预实验是你计划做的实验的一个小范围版本。通过这个实验，可以消除真正实验之前的任何问题。因为你需要的不是报告实验结果，而是你可以在与预实验中打破一些实验的规则。例如，你可以用甜言蜜语说服你的朋友来充当实验被试，甚至你自己也可以作为一个被试。你也可以中途改变自变量水平、停止实验或者只做一部分实验，这些都取决于你在预实验过程中了解到了的内容。



预实验中的被试

在做预实验时，你有时候会发现，写在纸上的设计看起来很好，但不一定可行。例如，有一次，我在做预实验时发现，我设计的一个看起来简单的实验竟然至少需要3个实验者进行操作。预实验也可以帮助你确定自变量的水平是否是符合你的期望。一个在计划阶段看起来可行的实验，可能在实验室里就无法实现。通过尝试，你可以改变一个明显不合适的自变量范围，避免投入大量的时间和精力。预实验对将来的实验具有导向作用，它能够指导实验者通过未经标示的水域。

虽然文献检索和预实验可以为你选择合适的自变量范围提供良好手段，但是，最后，你自己还是不得不进行猜测。如果你被证明是正确的，那么，你可以宣告自己的判断正确。如果你错了，你只能说自己运气不佳。

选择因变量

我们从第2章中了解到，因变量是对行为的测量。我们可以选择有限的行为进行测量。因此，在选择因变量时，我们必须确定将要测的内容。

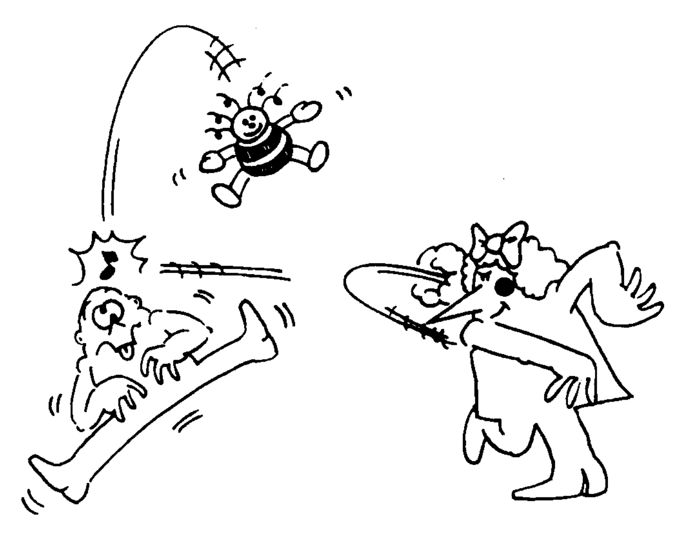
操作性定义

让我们回到"暴力电视演出会导致孩子攻击性的改变吗？"这个问题上来。在这个实验里面，我们就是要测量攻击性，但是，我们需要对攻击性进行操作性定义，以便我们能够确定孩子在观看了暴力电视节目后行为是否发生改变。

在这个实验里，形成一种操作性定义的方式是请一组评判者看电影，电影中每一个孩子都在自由玩耍。然后，用7点量表评价孩子的攻击性。或者我们可以告诉每一个孩子几个关于孩子在挫折环境下的故事，问孩子在相同的环境下会如何做。接下来可以使用"直接—攻击"反应的数量作为攻击性的测量指标。另外一个方法是观察孩子们选择玩具时候的表现。例如，我们先将玩具分成攻击性玩具（枪、坦克和刀子等）和非攻击性玩具（卡车、工具和布娃娃等）。然后，测量每一个孩子玩每种玩具的时间占总时间的百分比。你会想出许多行为，这些行为都对孩子的攻击性有着预测作用。

有时候，尽管一个因变量看起来很直接，但对它进行操作性定义也存在问题。例如，两个研究者想知道进化心理学理论的预测是否能够得到杀人案例的支持

① 我想pilot的意思是"引向未知的地方"，就像一个轮船驾驶员来到驾驶舱转动方向舵，驶向未知水域。



这是一个非暴力男孩吗?

(Daly & Wilson, 1988)。这个理论预测,人们不可能杀害与自己有血缘关系的一起生活的人,而更可能杀害与自己生活在一起的无任何血缘关系的人。现如今,从一个特定样本中找出杀人犯是一个非常简单的事情。但是,杀人到底是什么样的呢?在好多个国家,杀人包括所有的"谋杀、谋杀未遂以及过失杀人"。在这个研究中,谋杀未遂的和过失杀人都应该包括在内吗?对大多数过失杀人者而言,如粗心大意的车祸,他们并没有杀人动机。杀人意图很重要吗?如果意图重要的话,也许谋杀未遂者应该以杀人犯论处。

难道只有在法律上证明有罪的案件才可以被认为是谋杀吗?起初这个结论看起来很正确,我们不希望包括被告是无辜的案例。但是,计算有罪的人数可能更加误导人。在某一年底特律被判谋杀的案例中,20 个男人被指控杀死了他们的妻子,9 个女人被指控杀死了她们的丈夫。你可能得出结论男人杀妻发生得更频繁。然而,实际上,女人杀配偶更频繁。但是,杀夫案中有 75% 被驳回不予审理,而杀妻案则只有 20% 被驳回。正如研究者所指出,计算那些在法律上被证明有罪的案子的数量或许涉及更多的是刑事法庭的行为,而不是罪犯的罪犯行为!不幸的是,就像这个例子描述的那样,对因变量的操作性定义与对自变量的操作性定义一样难。

就因变量而言,我们不仅必须对它进行操作性定义,而且还不得不了解测量是否是可信的和有效的。

信度和效度

如果我们在相对平稳的条件下,重复测量很多次,都得到了相同的结果,那么,我们可以说测量装置是完美可信的。结果变化越大,测量装置的可信度越小。例如,一把橡皮尺子就可能不是非常可信的。某一次用它测量桌面长度时是 18 英寸,而下次测量时变成了 31 英寸。为了找出尺子的可信性,我们必须对很多物体至少测量两次。然后,计算结果的相关性(参看第 1 章)。如果第一次测量的结果与第二次很相似,相关度很高,那么,我们可以假设测量工具是可以信赖的。如果相关性很低,我们就知道测量工具不太可靠。

继续使用暴力电视节目的例子。我们可以让两组评判员对每个孩子相同的行为视频进行评价，然后，比较两组评判员对孩子攻击性的评分。如果两组评判员给出相似的评分，我们就可以认为，这些评判员的评分是可信的。

当因变量是成就测验、态度测验或人格特质测验等分数时，通过正式的方法估计信度（relia bility）是非常重要的。一个标准化测验的信度应该是已经被检验过，这种信度的统计值应该写在测验手册中。然而，如果你自己编织测验或问卷，那么，你必须自己确定其信度。确定信度的方法有以下几种。最简单的方法是重测信度（test-retest reliability），即采用同一测验对同一组被试进行两次测试。信度由两次测试分数之间的相关系数决定（参看附录 A）。然而，给同一个人的第二次测试结果可能受到第一次测试的影响。两次测试实施之间发生的事件也可能影响测试结果。因此，确定信度的第二种方式是复本测验（alternative-form）法。在这种方法中，第二次测验和第一次测验的项目相似，两次测验的被试也相同。也是计算每个人的两次测试结果之间的相关。确定信度的第三种方法是使用分半信度测验技术（split-half）。这种方法是将一个测验分数在统计上分成两半（例如，奇数项和偶数项）。然后，计算这两半分数的相关。使用每一种技术的优缺点见表 7-1。如果你的因变量不是一个测验分数，你可能不需要通过正式方法确定它的信度。不过，你应该意识到信度测量的必要性。

表 7-1　确定测验信度三种方法的优缺点

信度方法	优　点	缺　点
重测法	使用相同的测验方法。 简单易行。	第一次测验可能影响第二次测验。 反应会随时间而变化。
复本测验法	减小了重复项目的影响。 重测时间间隔小。 适用于前—后测设计。	使用不同的项目降低了信度。
分半法	减小了重复项目的影响。 没有时间间隔。 一次就可以完成。	使用不同的项目降低了信度。 测验项目较多。

效度（validity）[1]用来确定我们测量到的东西是否是我们想要测量的东西。假设你有一个 12 英寸长的木尺，但它的实际长度是 24 英寸，因为这个尺子上的每一英寸实际上都是两英寸。这种情况下，我们可以多次测量桌面的长度都是 11 英寸。这个工具虽然可靠，但测量结果是错误的，因为我们每次测量的结果都与实际不相符。因此，我们也需要知道测量工具是否有效，也就是说，它们的测量结果与一个已知有效的标准测量工具之间是否一致。

例如，在对攻击性进行操作性定义时，假设我们决定测量每个孩子花在攻击性和非攻击性玩具上的时间的百分比。如果我们的秒表工作正常，那么，这个测量结果也许是可信的，因为当我们再次测量这个行为时间时，我们可以得到相同的读数。然而，人们可能会说我们测量的攻击性是无效的。他们也许会认为，孩子们倾向于

① 对各种效度的详细讨论见第 2 章。

玩弄那些自己会玩的玩具。因为他们在暴力电视节目上见过枪、坦克和刀子,他们就选择玩这些玩具。他们也可能认为,孩子们既可以以攻击性方式使用坦克、工具和洋娃娃,也可以以非攻击性方式使用它们。为了说服别人你的测量是有效的,你可以将它与一些你们两个都认为是攻击性的有效测量的标准相比较,如果你的测量工具和标准相一致,那么,你就可以认为这是一个有效的测量工具。

当将一个测验分数作为因变量时,采用正规的方法检验测验的信度和效度是非常必要的。最弱的效度形式是表面效度(face validity),它是指表面上看起来测验的内容好像是应该测量的内容。显然,表面效度比较主观,以至于它在科学研究中没有得到广泛使用;所有的研究者都认为,自己的测验具有很高的表面效度。一种更加正式的有效程序是建立内容效度(content validity)。它是按照内容对测验涉及的主题进行仔细而详尽地分析。然后,从每一个确定的内容中取出具有代表性的样题。例如,如果你想做一个评价阅读理解能力的测试,那么,测试项目应该包括涉及该章中每一个主要概念,如内容效度。第三种有效性程序是建立预测效度(predictive validity),即确定测验能否成功预测某些特定标准。例如,高中学生进入大学的入学考试成绩在一定程度上可以预测他们在大学中的成绩(GPA)。测验成绩和GPA之间的高相关系数就说明测验具有较高的预测效度。同时效度(concurrent validity)也是将测验成绩与一个标准进行比较,在这种情况下,两个测试应该同时进行。例如,如果我们想要设计一个问卷,通过父母填写,以测量他们孩子的攻击性。我们就可以通过计算每个孩子家长填写的问卷分数和老师的攻击性评分之间的相关,确定它的同时效度。由此可见,测量一个因变量的效度可能比测量它的信度更难。因此,我们能做的最好的事情就是使测验不要在逻辑上出现任何漏洞。

可直接观测的因变量

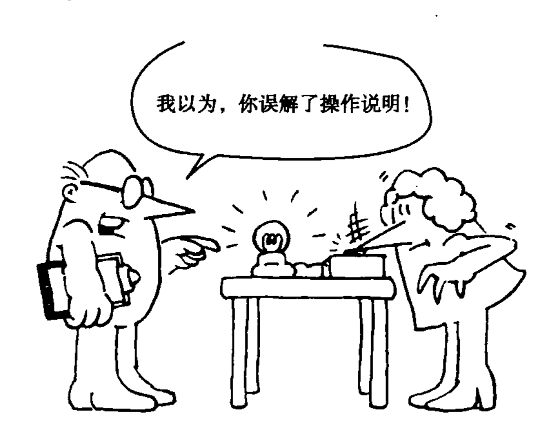
你越是直接观测一个行为,你的测量结果引起的争论就越少。然而,如果你的兴趣是决定人的心理过程,那么,你就应该认识到所有因变量在某种程度上都是间接的。例如,假设你对记忆感兴趣,想要比较将记忆材料的两种表征方式。一周后,测量被试记住了多少。那么,你该测量什么呢?

显然,就去问他们记住了哪些内容。但是,假设他们无法回想起两种方式呈现的任何材料。你能得出结论说他们什么也没记住吗?你或许可以给他们一个再认测验,让他们区分哪些是先前呈现的材料,哪些是新出现的内容;然后,计算他们的正确率。或者你可以让他们重学材料,测量第二次学习以前学过相同的材料所节约的时间。两种方法可能给出不同答案:人们记住了多少?我希望你能从这个例子看出,因变量(甚至那些初看起来可以直接观察的变量)与你感兴趣的行为之间的联系也只是间接的。

单一因变量

假设我们想知道,让人们对出现信号进行按键反应时,人们对亮光的反应比对暗光反应更快。我们可能在光亮出现时启动一个时钟,而当被试按按键时,停止该时钟。我们必须要看到,这里仅测量了反应的一个特性。我们本来可以选择反应的其他任何特点——如人们如何按键。在一次试验中被试是否会将手指移到键的一

侧，而在下一次反应中直接按在键上呢？在一次试验中，她的第一次尝试是否错过了时机？而另一次试验中，她的第一次敲击是否太轻，后来才使劲按下去了呢？从诸如此类的反应中，我们只选择测量反应的一个特性，即从亮光出现到按键。换句话说，我们选择的是单一因变量。



我们选择的任何单一因变量都不一定是最合适的测量指标。例如，假设我们让被试者使用铅笔画出镜子里的五角星的轮廓。因为镜子里的东西都是反的，大多数人在前几次试验时，都会觉得这个任务很难。如果我们想测量这个任务从第 1 次试验到第 10 次试验的进步程度，那么，什么因变量能够最好地反映这种进步呢？这类测验中标准的因变量是被试者沿五角星画线时所用的画线次数。图 7-1 演示了两名假想的被试者的画线情况，我们姑且称之为被试者 1 和被试者 2。在第 1 次试验中，被试者 1 画出边界 20 次，而在第 10 个试验中只画出了 6 次。对这个被试者来说，因变量反映了行为上预期的进步。但是，被试者 2 的两次测试都画出了五星轮廓边界 14 次。因此，因变量显示，被试 2 在画镜像五星任务中没有进步。你相信这个结论吗？

这里的基本问题是，即使使用一个直接的可观测的因变量，如边界交叉数目，我们也必须关注它的效度。画出边界的行为只是镜像测量指标之一。它是一个有效的测量指标吗？其他因变量可能更好地反应镜像画线行为。另一种选择是，我们测量画线的总长度，计算落入五角星轮廓边界内的百分比；我们也可以测量每一次试验中边界与画线之间的区域；或者记录被试的所用时间，找出他们在第 10 次画线时是否变得更快。

多重因变量

在实验中，使用多重因变量（multiple dependent variables）可以增加测量适当行为的可能性。实际上，在实验心理学的某些领域，人们认为，只报告一个因变量是非常不合适的。例如，很多类型的研究使用选择反应时（choice reaction time）作为因变量。选择反应时是指当几个刺激中的一个出现时，对它反应所需要的时间①。显然，如果人们想要尽可能少的犯错误，他们必须反应得更慢些。如果他们愿意出错，

---

① 由于前面没有使用刺激（stimuli）一词，在这里有必要说明，stimulus 是单数，stimuli 是复数。因此，注意这里的用法："this stimulus is（刺激是），this datum is（数据是）；these stimuli are（刺激是），these data are（数据是）。"

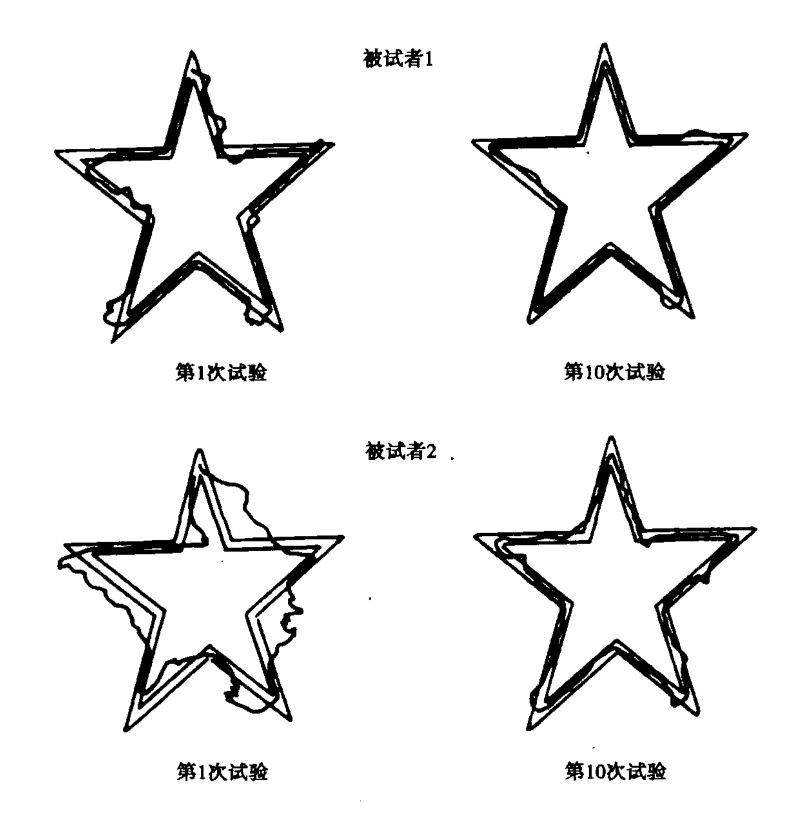


图 7-1 两个被试在第一次实验中的五角星轨迹追踪结果

则可以反应得更快。这种速度—准确性之间的权衡使得速度和准确性都应该被作为因变量报告出来。如果我们对行为的所有方面感兴趣,只有一种测量指标是不够的。因此,较好的杂志不会接受只报告速度或者只有反应正确率的文章。

复合因变量

虽然尽可能多地报告行为的不同方面通常是个好主意,但是,这样做使得结果解释变得更加困难。假设我们有 4 个因变量:一个测量指标表明,不同条件下的成绩大幅度提高;另外两指标没有变化;最后一个指标则有些轻微的降低。为了说明行为的整体变化,我们需要将单个因变量组合起来,构成复合因变量,这种复合因变量可以说明行为的总体变化情况。

在实验心理学的许多领域中,如智力测验,都使用复合因变量。用韦氏成人智力量表测 IQ 的一般智力测验就是复合因变量的一个实例。IQ 由两个分量表组成,一个语言分量表和一个操作分量表。每一个分量表的得分都来自于这些分测验。例如,语言得分来自于一般信息、数字广度、词汇、数学、理解和熟悉程度等分测验。智力测验的基本理念是使用一个能够体现智力特点的单一测量指标是可行的。当然,不是所有心理学家都同意单一数字能够充分代表智力,但是,在心理学测验中使

用复合因变量已经形成了传统①。

第二种类型的复合因变量由一个单一测量指标的几个方面组合而成。这些方面在不同时间或不同情景下都可以使用。节省百分率就是用在记忆研究中的这样一个因变量。例如，假设一组被试在小时候学会了骑自行车之后，直到他们 40 岁时都没再碰过自行车。我们可以让他们再学骑自行车，通过几次练习，直到他们能够在自行车上骑 1 分钟以上。假设他们需要 7 次才能做到这一点。我们可以把这个数字和另外一组 40 岁却从来没有接触过自行车的被试进行比较，让他们也练习，直到骑在自行车上达 1 分钟。假设这组平均需要 14 次。接下来，我们就可以计算早年学过骑自行车的被试再一次学会骑自行车节省的百分率：

$$\text{节省百分比} = \frac{\text{初学次数} - \text{从学次数}}{\text{初学次数}} \times 100\%$$

在我们的例子中，应该是：

$$\text{节省百分比} = \frac{14-7}{14} \times 100\% = 50\%$$

虽然这是一种复合因变量，但你可以使用一个简单的数字表示由自变量（过去有过骑自行车的经历）引起的变化效应。

你也可能还不是很清楚，这些复合因变量是由何而来，或者为什么它们是合适的测量指标。但是，如果你在心理学的某些领域做研究，那么，你可能对其他复合变量非常熟悉。你甚至会发现，有一天，自己也开始设计复合因变量了。

间接因变量

有时候，直接观察你感兴趣的行为是不可能的。然而，我们知道，可重复、可观察和可检验的科学测试要求我们研究的行为是可被公众观察的。那么，我们在诸如情绪、学习或者智力等领域如何开展科学研究呢？我们需要选择一个间接变量，这些变量随着我们感兴趣的内部行为的改变而变化。

生理测量指标

最流行的间接测量指标可能就是生理指标，生理指标所基于的观点是，如果行为是一种特有事件，如情绪；那么，身体的生理指标也会随着这一事件的变化而发生改变。由于现代技术可以帮助我们观察身体的生理指标的变化，实验者就可以利用这些变化观察特有事件发生的过程。

当然，当我们用生理学指标推论内部状态时，我们是在假设一个特异性的生理模式能够准确地反应一种内部状态。例如，多导生理记录仪或测谎仪可以测量 4 项

---

① Stephen Jay Gould (1981) 在他的著作《有关人的错误测量指标》(*The Mismeasure of Man*) 中对于使用单个数字作为像 IQ 等的指标提出了强烈批评。他认为，这是上个世纪描述在人的价值方面最主要的科学上的误用，并相信这种复合因变量最多不过可以用来显示社会等级和差异。

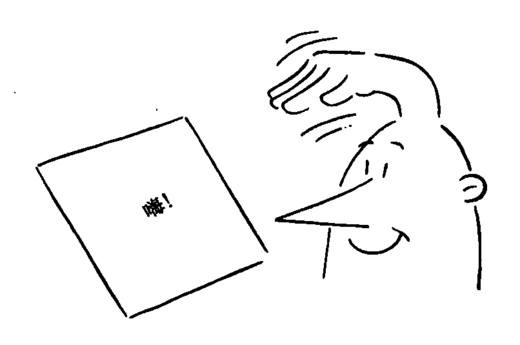
生理指标,包括呼吸频率、心率、血压和皮肤电等①。研究者通过这些测量指标确定一个人是否在说谎。有人怀疑使用生理测量指标背后的假设是否正确。因此,在大多数法庭上,只有原告和被告都同意使用测谎仪时,它的结果才可以被作为证据。而且,最近,联邦法律已经对聘用筛选中使用多导仪测试进行了严格限制。

当研究者声称,这些测量指标可以作为一些情绪状态的标记时,其他生理测量指标就逐渐被人们重视起来。然而,这些测量指标并不像其他研究者展现的那样,同一种生理指标会随着不同的内部过程而发生改变。例如,一个叫 Hess 的研究者一度声称,一个人在想高兴的事情时瞳孔直径会增加,想不高兴事情时会缩小。有那么一阵子,麦迪逊大道的广告巨头对此深信不疑,他们就使用瞳孔反应选择杂志广告。其他研究者已经发现,瞳孔的直径也许能够更好地说明人的加工信息数量,而不是感受到的情绪(Johnson, 1971)。于是,瞳孔测量专家们再也不像以前那样受到麦迪逊大道的欢迎了。

最近,一些研究者报告说,一个人声音的特点可以用作"心理应激评价"。通过磁带记录一个声音,慢速播放,测量声音频率的某些特性,这些研究者认为,他们能够辨别出那些处于紧张状态下的人,如说谎的时候。这些说法并没有得到研究支持,很多研究者认为,这种测量指标并没有什么价值。

在过去的几十年中,发展最快的心理学领域是脑成像。在成像过程中,被试完成各种不同任务,同时记录他们的大脑活动。早期的研究中通常以脑电图(EEG)为大脑活动的一般指标。然而,这种一般的活动模式除了对确定一个人的整体唤醒水平外并没有什么用处。最近,研究者可以重复呈现一种刺激,然后,以刺激呈现时间或反应时间为起始点,对脑电波活性进行叠加。通过事件相关电位(ERPs),可以对脑电图的波峰和波谷上的微小变化进行分析,以判断人们对刺激的反应或任务需要的认知过程。

就在更近,研究者采用生理学技术能够将被试执行任务时大脑内部的活动进行定位。使用最为广泛的技术是功能磁共振成像(fMRI)。研究者使用 fMRI 技术测量血流到脑内各个区域的情况。这种方法的一般逻辑是,当某个特定脑区进行信息加工时,这种心理活动需要该区域的放电神经元增加。随着神经元不断放电,它们需要的血液也增加。因此,如果研究者要求被试完成一个任务,并发现他们大脑的某个区域血流增加,那么,就可以推论,该脑区负责完成这项任务。



有些刺激产生一种特定的脑电波

例如,假设我用 fMRI 测量你在看某个单词时大脑的血流量。一种条件下要求

① 皮肤电反应不是由于手中的物品太多引起,它是衡量皮肤传到一个微弱电流能力的指标。虽然在技术上它并不精确,使用它的逻辑类似于:因为湿皮肤比干燥皮肤的导电性更好,当一个人"出汗"时的皮肤电与"冷静和安静"时的皮肤电就会有所不同。

你读这个词;另一种条件下对词义作判断。在每种条件下,我都能画出一张图表明脑内各区域血流量的图。通过从第2张图中减去第1张图,我们可以推论,哪些区域参与了词义加工。研究者不仅使用fMRI开展这类工作,还利用计算机断层扫描(computerized axial tomography, CAT scans)、正电子发射断层扫描(PET scans)和多导脑电图等技术。这些技术使人们在理解人脑功能方面取得了巨大进展①。随着我们对这些测量指标的了解,使用生理指标的数量也肯定会增加。

行为测量指标

有些行为测量指标也可以被用作间接因变量。就像生理测量指标一样,人们行为方式的改变也能够反映人的内部状态的变化。

间接的行为测量指标在认知心理学的某些领域尤其重要。认知研究者对人脑这个"黑匣子"里发生的事情非常感兴趣,如阅读和问题解决。由于他们所面对的是这个盒子的输入(刺激)和输出(反应),因此,他们就不得不寻找聪明的方式推论盒子里面到底在发生什么。例如,假设我们想知道在完成某个工作时,需要加工多少信息。如果我们假设大脑在加工认知信息时提供的资源有限,那么,确定有多少信息被加工的一种方式是测量一个反应所需的时间。信息加工得越多,反应的时间就会越长。然而,反应时只是整个任务的一个单一测量指标,而无法给出次级任务的加工需求,如编码或反应选择。

双任务方法(dual-task methodology)是一种间接方式,它可以判定执行任务时的加工需求。这种情况下,当执行第一个任务时,第二个任务也会出现。被试要完成第一个任务,同时尽可能利用剩下资源完成第二个任务。通过测量第二个任务的行为指标,就可以推论完成第一个任务的加工需求是什么。在第二个任务上的表现越好,第一个任务所需的资源就越少。例如,第一个任务可能是读一个句子。而让被试读句子的时候,呈现纯音,要求被试无论什么时候听到纯音都尽快按键。我们可以推论,对纯音的反应越慢,加工句子的认知需求就越多。试过几次后,就可以画出在读句子时,对纯音反应时的曲线,于是,便可以得到了句子加工所需的资源轮廓图(Martin & Kelly, 1974)。

与所有间接行为测量指标类似,这种测量指标也只能像它所假设的那样。在双任务方法中,主要的假设是,所有认知任务所需的资源都来自一个单一的加工资源库。有些研究者对这种基础假设提出了质疑(Navon & Gopher, 1979; Wickens, 1984)。我们现在的确有证据支持存在多个资源库,此外,资源库的使用取决于这个任务是视觉,还是听觉的;是空间的,还是语义的等(Wickens, 1984)。虽然双任务方法的一些假设仍然存在激烈争论,但在很多情况下,这项技术为加工资源研究提供了非常好的测量指标,并且被广泛应用。

其他间接行为测量指标与双任务方法的假设不尽相同。然而,一般而言,测量指标越间接,对假设的解释就越要精细,而对推论的信心就越弱。间接测量指标的优点在于,它提供了一种无需直接测量就可以研究实验问题的方式。只要我们在使

---

① 然而,有研究者提醒人们,在探索人脑奥秘的过程中,不能毫不怀疑地全盘接受脑成像结果(Van Orden & Paap, 1997)。

用这些间接测量指标时，充分了解假设，它们就是非常有价值的测量工具，可以帮助我们揭示未知事件的本质。

小　结

在实验中选择自变量时，首先，你必须对变量进行操作性定义，以便其他实验者在相似实验中可以仔细分析相同的操作。自变量水平的选择也非常重要，选择范围要足以出现实验效应，但也要考虑到现实条件的限制。一组试验或预实验有时候会帮你选择自变量。

因变量也必须有操作性定义。此外，我们还必须能够保证因变量是可信的和有效的。如果每次测量都会得到相同的结果，则它是可信的。当使用测验分数作为因变量时，测验的信度可以由这样几种方式决定，即重测、复本和分半。如果因变量与被广泛接受的标准一致，则它是有效的。建立测验效度的方法有表面效度、内容效度、预测效度和同时效度。直接可观测的因变量相对易于测量，但有时单一因变量难以确定。有些研究领域要求报告多重因变量，或者将因变量组合成一个复合因变量。当我们感兴趣的行为不是可以公开观测的时候，我们就使用间接因变量。生理测量指标可以反映内部状态，但经常很难解释。行为测量指标，如双任务方法，也可以用来确定被试内部状态。

8

如何选择“被试间设计”和“被试内设计”

也许世上只有两种人，一种人习惯将人分成两类，而另一种人则不做这种区分。

ROBERT BENCHLEY

现在你已经选择了一个可操控的自变量和一个可测量的因变量。如果人与人之间非常相似,那么,只需一个被试去做你的实验就可以了。但事实上,人并不相同,人与人之间的差异构成一个丰富多彩的世界;但这对于一个实验者来说,则是不幸的。由于我们每个人都是不同的,因此,你必须使用被试样本,并通过统计手段(如计算平均)降低来自被试方面的变异。然而,在降低被试引起变异的方法上还有多种选择,这依赖于你如何将被试分配到自变量的不同水平上。

被试分配方法有以下两种,即让每个被试只接受自变量一个水平的处理,或者接受所有处理水平。第一种方法由于自变量至少是在两个被试①之间进行操控的,因此,通常被称为"被试间设计"。第二种方法由于自变量是在一个被试身上操控的,通常被称为"被试内设计"②。表8-1显示了在含有两个水平的单个自变量实验中,这两种方法是如何分配被试的。在表8-1上半部分的设计中,2个分别由10人构成的被试组被分配到了不同水平上;在表8-1下半部分的设计中,10个被试被分配到了两种水平上。

表8-1 被试间设计和被试内设计中的被试分配方式

组间	自变量	
	水平1	水平2
	被试1	被试11
	被试2	被试12
	⋮	⋮
	被试10	被试20
组内	自变量	
	水平1	水平2
	被试1	被试1
	被试2	被试2
	⋮	⋮
	被试10	被试10

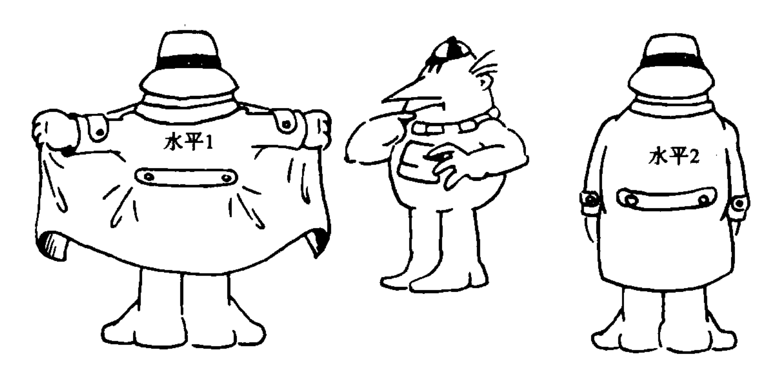
假设我们想做一个休息是否会提高学生成绩的实验,一种实验条件是让学生连续2个小时学习某种材料;另一种实验条件是让学生总共学习2小时,但每半小时会让学生休息5分钟。每种条件结束后,都对学生进行测验。我们可以使用被试间设计,将随机选择的学生所构成的不同实验组分配到每种实验条件中;或者使用被试内设计,此时同一个实验组要在每种实验条件下学习不同的材料。如果在两个实验组中使用不同被试,则不仅存在实验组内的被试差异,还存在实验组之间的整体差异。如果我们使用同一组学生,尽管学生之间在学习能力上存在个体差异,但实验组之间是没有整体差异的,因为他们是同一组被试。然而,由于此时我们必须使

---

① 本书至此,作者沿用美国心理学会(APA)的风格,将参加实验的人称为participant,而不是subject。然而,由于普遍使用的实验设计的名称、统计方法的名称都没有与APA手册上的名称一致起来,在本章中完全使用"participant"这个术语会引起一些不便。因此,我仍旧使用大家认可的名称(如被试内设计"within-subject")来表示实验设计类型和统计检验方法,但我会继续将参加实验的人称为participant。我希望在不久的将来,大家能有一个统一的术语。

② 有人将"被试内设计"称为"实验处理×被试设计(Treatment × Subject design)",或重复测量设计(repeated-measures design)。被试间设计有时称为独立组设计(separate groups, independent groups design)。

用不同的学习材料,那么,材料之间在难度上又会存在差异。下面我们将更详细地讨论一下这两种实验设计的优缺点。



组间设计:每个被试只接受一个水平的处理。

被试间设计

优 点

被试间设计最大的优点是被试接受自变量一个水平的处理不会影响到他们在其他水平上的行为反应。因为每个被试只接受了一种水平的处理,对某一个被试而言,完全可以有效忽略其他水平的影响。

在前面曾介绍过一个我和我学生做的实验。为了考察"可获得性"启发式,我们让被试分别列举核能的三个优点,或者三个缺点,或者三个优点与三个缺点,列举完毕后,再让被试通过等级评价表明他们对核能的支持程度。这个实验使用的是被试间设计,因此,每个被试要么列举三个优点,要么列举三个缺点,要么列举三个优点和三个缺点。如果使用被试内设计,结果会怎样呢?这种情况下,每个被试需要先列举三个优点并评价对核能的支持程度;然后,再列举三个缺点并评价支持度;最后列举三个优点和三个缺点并评价。这种被试内设计能否得到我们想要的结论呢?请记住,我们的假设是对核能的支持程度会受到列举优点还是缺点的影响,这背后的逻辑关系是,通过列举,那些被列举的理由会变得更容易被获取到——在该被试大脑中更容易被通达。但是,某个优点一旦变得更容易获得,那么,要多久才能使它变得不容易被获取呢?事实上,如果我们使用了被试内设计的话,一旦被试接受了前两种处理条件,即先列举了三个优点又列举了三个缺点,这也使他们已经接受了第三种条件的处理,即列举优缺点各三个。

对于许多实验来说,比如刚才描述的实验,在逻辑上它是不可能使用被试内设计的。因为我们不能消除先前的处理对其他处理水平的影响。尽管有些情况下,在逻辑上可以消除这种影响,但如本章后面所讨论的那样,被试内设计会变得更加复杂。结果是,即使进行了非常周密的设计,有时我们也不能十分确定是否已经彻底消除了先前处理的影响。而被试间设计不会有这样的问题,因此,有时会首选这种实验设计方式。

被试间实验设计在实践上也有一些优势。由于被试只需要完成自变量一种水

平下的处理，在此水平上，我们可以收集到更多的数据。由于被试不太可能疲劳或失去实验兴趣，对每个被试而言，很容易保证较短的实验时间。被试间设计还可以避免将被试带入到其他水平的处理中，虽然被试内设计可以使被试的数量大大减少。

缺点

被试间设计最大的缺点就是，被分配到自变量不同水平的实验组可能会在某个维度上不对等，这个不对等的维度可能会导致所要测量的行为发生偏向。只要实验组是由不同人构成的，那么，它就有可能存在巨大差异。例如，在观看暴力电视节目是否会引起儿童攻击行为的实验中，可能会出现这样的情况：所有被分配到观看暴力电视的儿童都来自于有功能缺陷的家庭，如有吸毒史等；而没有观看暴力电视的儿童均来自于健康家庭。如果将儿童随机地分配到不同的实验组中，上述情况就不会发生。

当使用被试间设计时，通常采用随机方式分配被试。这种分配可以通过多种方法实现，例如，给被试编号、掷硬币，或者从如附录 C 随机数表中进行选取。心理学实验经验较少或统计经验较少的研究者似乎对随机选择的过程有些缺乏信心。他们常常认为，随机就等同于杂乱无章和轻率；他们还认为，即使使用大规模的实验组，实验组的行为反应仍可能会存在巨大差异。随着实验经验的丰富以及对统计取样过程的充分理解，研究者会更有信心地使用随机分配被试的方法。另外，尽管随机化看起来缺乏秩序，但它至少是无偏向的，可使你以一种无偏向的方式分配被试。尤其是对于被试容量较大的实验组来说，实验组之间在行为反应维度上存在巨大差异的可能性相当小。最重要的是，数据分析的统计方法充分考虑了由随机分配造成的潜在差异问题。因此，在消除实验组之间的潜在偏向上，被试间设计中随机分配被试的方法是非常有效的。

被试内设计

虽然对于所有的实验来说，被试内设计并不是绝佳的选择，但它确实具有一些优点。

实践上的优点

被试内设计在应用中的一个优点很容易从表 8-1 中看出来，即实验需要的被试较少。在使用被试内设计时，如果任何处理水平都有足够数据点，且共需要 N 个被试①，那么，在两水平的被试间设计上则需要 2N 个被试，在三水平的被试间设计上则需要 3N 个被试，其余以此类推。

很多情况下，被试数量的增加也同时增加了完成实验的总时间。例如，在接受正式实验处理前，通常需要事先训练被试完成某种任务，那么，单因素两水平的被试间设计实验，训练被试所需的时间会两倍于被试内设计的实验。假设你想知道让人

① 在此使用 N 表示特定实验中任意数量的被试，如 10 个或 20 个。

记住一定数量的单词是否会与完成某个复杂的轨迹追踪任务有关，再假设学会追踪任务要用数小时。如果你增加自变量水平（如需要记住单词的数量），那么，采用被试内设计不会增加训练被试的时间，而在被试间设计中，由于增加了被试的数量，因此，也就导致总的训练时间增加。

总体而言，实验前被试都要完成几次练习试验，由于被试数量增加，练习的总时间也会增加。这些练习试验还用来减少"热身效应"——即随着被试逐渐进入实验准备状态，在前几个试验上通常会出现反应速度的迅速提升。

除了被试数量大以外，被试间设计的另一个不足之处是，有时能参加实验的被试数量有限，当被试需要满足某些条件时，尤为如此。例如，实验需要的被试是飞行员、赛车手或芭蕾舞演员，或者是精神异常、色盲或左撇子等①。遇到这种情况时，被试间设计可能找不到足够能满足实验要求的被试，那么，此时就必须要使用被试内设计。

统计上的优点

除了效率上的优势外，被试内设计在统计上也很有优势。我们可以简略地看一下第 12 章中的统计方法，在此我先提及几个概念。

在推论统计中，实验者试图从自变量不同水平上得到的结果去推断数据的差异是由行为反应上真实存在的差异造成的，还是由随机误差造成。为了得出推论，在多数统计检验中，实验者需要将在两个水平上平均反应之间的差异与每个水平内部误差的估计值进行比较。通过检验，如果不同水平之间的差异非常大或同一水平内部的变异非常小，那么，实验者可以肯定地说差异是客观存在的。下面用一个例子对这一原则的逻辑关系进行详细说明。

假设一个跑鞋制造商想知道，应该卖给男子百米短跑运动员钉长 7 mm 的鞋呢，还是钉长 13 mm 的鞋。为了测验两款鞋子，制造商可以从某大学中随机选择 10 人穿其中一款跑鞋，在同一大学再选另外 10 人穿另一款跑鞋，这两组人的短跑时间会存在差异，从 300 磅体重、38 岁的调酒师到 125 磅体重的 19 岁的中卫球员等。他们的成绩如表 8-2 显示的那样。如果你计算两组的平均值②，你会发现，穿钉长 7 mm 跑鞋组比钉长 13 mm 组快 0.5 秒。这个差异能否使你相信短钉鞋更适于百米短跑呢？

假设制造商决定做第二个实验，这次选择男子短跑队员做被试，将他们随机分配到 7 mm 组和 13 mm 组，得到的结果可能像表 8-3 所示，这次又发现了穿短钉鞋会产生 0.5 秒钟的优势了，那么，这时的数据能否使你相信穿短钉鞋更好呢？

毫无疑问，你更愿意相信第二个实验中的差异是真正的差异，因为第二个实验中的每个组内部的变异更小，你会觉得此时实验组之间的差异更不可能完全由随机误差产生。

---

① 有时被戏称为 lefties（This was just a sinister joke!）

② 第 12 章和附录 A 中会详述，平均值是被试得分之和除以被试的数量。

表 8-2 两实验组中每个被试的百米短跑时间

穿 7 mm 钉长跑鞋的男子	时间(秒)	穿 13 mm 钉长跑鞋的男子	时间(秒)
Mike	11.7	Don	15.7
Bob	18.2	Hector	13.4
Homer	12.2	Ron	18.0
George	15.4	Tom	12.8
Harry	15.8	Steve	13.6
Gordon	13.2	Dale	19.0
John	13.7	Pete	16.2
Bill	19.1	Juan	11.9
Randy	12.9	Dan	14.6
Tim	16.0	Paul	18.0

穿 7 mm 钉长跑鞋的男子的平均速度 = 14.82 秒;穿 13 mm 钉长跑鞋的男子的平均速度 = 15.32 秒。平均差 = 0.5 秒。

表 8-3 两运动员组每个被试的百米短跑时间

穿 7 mm 钉长跑鞋的男子	时间(秒)	穿 13 mm 钉长跑鞋的男子	时间(秒)
Art	10.6	Rod	10.8
Simon	10.3	Frank	11.0
Nick	10.3	Walt	10.8
Daryl	10.2	Gary	10.6
Ralph	10.4	Ken	10.8
Will	10.0	Bryan	10.7
Reuben	10.2	Dick	10.6
Ed	10.1	Stan	10.7
Fred	10.3	Rich	10.7
wayne	10.4	Mark	11.1

穿 7 mm 钉长跑鞋的男子的平均速度 = 10.28 秒;穿 13 mm 钉长跑鞋的男子的平均速度 = 10.78 秒。平均差 = 0.5 秒。

第一个实验中,被试短跑成绩的大部分变异是由于百米短跑中较大的个体差异引起的,而不是钉鞋引起的。第二个实验中,通过选择更为相似的短跑运动员,由个体差异引起的大部分变异被消除。

如何使两个组的被试更为相似呢?当然是使用相同的被试了,让部分被试先完成处理一,部分被试先完成处理二。你可能会发现,为什么仅有一个实验组的被试内设计具有统计优势了,它是将被试间的个体差异降到最小的绝佳方法。通过使用被试内设计,可以使你在主观上以及统计检验上更确信,自变量不同水平之间的反应差异是真实存在的差异①。

① 如果你有严谨的统计检验倾向,你或许对我这样深入浅出地介绍推论统计的做法嗤之以鼻,我在第 12 章中会更严谨些。

缺 点

既然使用被试内设计在实践上和统计上有那么多优点，为什么还需要使用被试间设计呢？不幸的是，被试内设计还有一些相当严重的缺陷。有些实验者甚至声称，这些缺陷使得被试内设计几乎毫无用处，尽管这样的观点很有争议。“总有一天，没有任何一个知名的心理学家会再单独使用被试内设计，而不与被试间设计结合在一起使用了，除非是将其用于某些特殊的目的”(Poulton,1973)。

正如讨论被试间设计的优点时那样，被试内设计的基本问题是，一旦被试接受了自变量某个水平的处理之后，不可能再将被试变为接受处理前的状态，接受实验处理会产生不可逆转的改变。因此，我们无法再将被试视为纯净的、未受污染的和纯真的被试。有些研究者认为，先前的实验处理导致被试产生了“迁移效应(carry-over effect)”。被试的改变还依赖于接受自变量不同处理顺序的影响，这种差异通常被称为“顺序效应(order effect)”。如果被试的行为反应受不同水平上的处理顺序影响，那么，被试内设计中就会产生处理顺序效应。

其中一种影响行为反应的顺序效应是学习，即在接受自变量某个水平的处理后，被试产生的学习会影响到接下来的行为。例如，假设我们想知道在标准的 QWERTY 键盘①上和在新设计的键盘上(使用频率较高的字母键在放松状态时手指的下方)哪个打字速度更快？由于在打字能力上存在着较大的个体差异，使用被试内设计的话，先选择 10 名被试，看他们需要练习多久才能在标准键盘上每分钟打 30 个词出来，然后，再让他们使用新设计的键盘，仍然看需要练习多久才能做到每分钟打 30 个词。假如被试平均需要练习 45 个小时才能在 QWERTY 键盘上达到上面的速度，而在新设计的键盘上只要 2 小时就可以了。那么，我们能否得出新键盘更容易的结论呢？显然不能。

在这个实验的前半部分，除学会使用 QWERTY 键盘外，被试还学会了普遍的打字能力，这种一般能力与特殊能力是相互融合的，当他们在新设计的键盘上打字时，这种一般的打字能力会毫无疑问地比实验刚开始时处于更高的水平上。因为首先使用 QWERTY 键盘，一般打字能力和特殊打字能力都要学习。因此，被试需要更长的时间才能学会。使用新设计的键盘经常发生在使用 QWERTY 键盘之后，由于已经学会了一般打字能力，因此，需要的学习时间较短。学习只是一种最常见的顺序效应，然而，还有其他的现象，如疲劳或成熟等。在实验过程中的任何时间，我们都需要注意可能发生的顺序效应，并且要保证自变量的作用不会受到呈现顺序的影响。

由于被试内设计的这些缺点，它在心理学的某些领域中很少使用。例如，学习和记忆的研究、社会心理学的某些领域(如态度形成)的研究，都希望被试会在某些

---

① QWERTY 键盘是根据标准键盘最上方的前 6 个字母命名的。研究表明，还有更优化的字母键排列方式会使打字的速度变得更快。然而，由于需要重新训练已经熟悉 QWERTY 系统的打字者，使得任何新的系统都不太可能被广泛接受。

特征上发生持久的改变,你不能告诉被试说,"好吧,现在请忘掉刚才让你记住的10个单词吧",或"请再变回到看这些宣传材料之前的态度上去吧"。在这些领域中,通过接受自变量某个水平的处理后,被试发生了不可恢复的改变。然而,在有些领域的研究中,先前处理的影响不大,如我们想研究人们区分两个纯音强度的能力,接受某个强度差别的处理后,就不可能会影响到他们区分其他声音差别的能力。这种情况下以及在其他领域(如感觉和知觉实验)中,被试内设计更受欢迎,使用频率更高。

平　衡

将学习这样的顺序效应降低至最小的方法之一就是平衡。从本质上讲,当你使用了平衡时,就意味着承认了存在潜在的顺序效应,也承认了不能控制它或者通过随机化将其排除,因此,要试着将额外变量的影响平均分配到自变量的各个水平上。通过这种方式,希望能够使顺序效应平衡掉,而不会造成自变量各个水平的反应产生偏向。

为了阐述平衡的概念,我用图8-1中的数值来说明。假如我们是全能的,已经知道了自变量和额外变量对行为反应的影响具体有多大。如果我们非常完美地做了一个单因素2水平(A水平、B水平)的实验,可能会得到图8-1中左半部分(模块1)中的结果。假设除自变量外,没有其他因素会影响到因变量,A水平对因变量的作用是1个单位,B水平是3个单位。因为要将这些量放到天平上,所以,我将数值转换为了重量大小,把重量放到天平上后,如右半部分(模块2)所示,我们发现,纯粹由自变量产生的作用是2个单位。

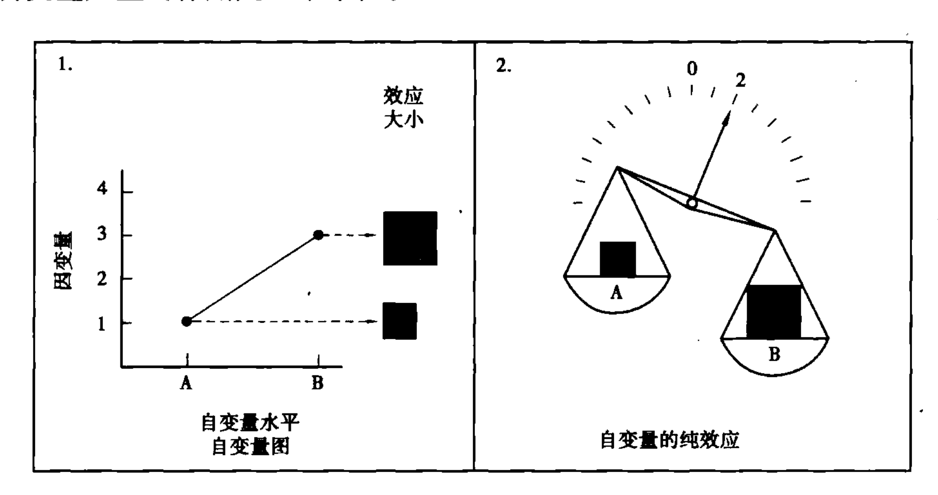
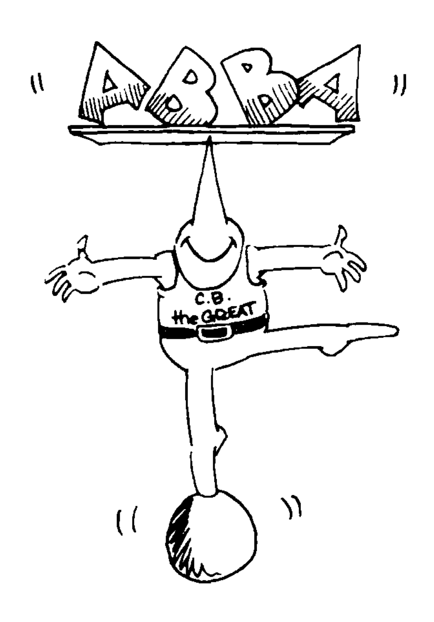


图8-1　图的左半部分自变量的A、B两个水平对因变量的效应,右半部分自变量的两个水平的纯的效应。

因为我们使用的是被试内设计,因此,被试不能在同一时刻接受两种水平的处理,显然,我们必须进行多次试验。假设存在某些额外变量,如学习,学习效应会随着每次试验而增加,如图8-2的左侧(模块1)所示。因此,在第1次试验中额外变量的效应是1个单位,但到了第4次试验时,就变成了4个单位。同样地,把学习效应转化为重量,我们希望把重量分配开,以使天平平衡。通过平衡法,当加入自变量



前，天平应该不会存在偏向。

一种非常常用的平衡模式被称为 ABBA 法，A 和 B 分别表示任意自变量的两种水平。ABBA 的顺序表示如何将每个水平分配到各个试验上。因此，在试验 1 中呈现的是 A 水平，试验 2 和试验 3 中呈现的是 B 水平，试验 4 中呈现的是 A 水平。每个被试都接受所有试验。

图 8-2 的右侧（模块 2）显示出，当试验 1 和试验 4 的重量放在天平的 A 侧，试验 2 和试验 3 的重量放在天平的 B 侧上时，两侧的重量相等。当我们再把表示自变量的阴影重量添加到天平上后，组合效应的净值是 2 个单位，与最初只有自变量时一样。因此，这种无偏向的结果正是我们想通过平衡方案所要达到的目标。

在你陶醉于平衡带来的好处时，需要提醒

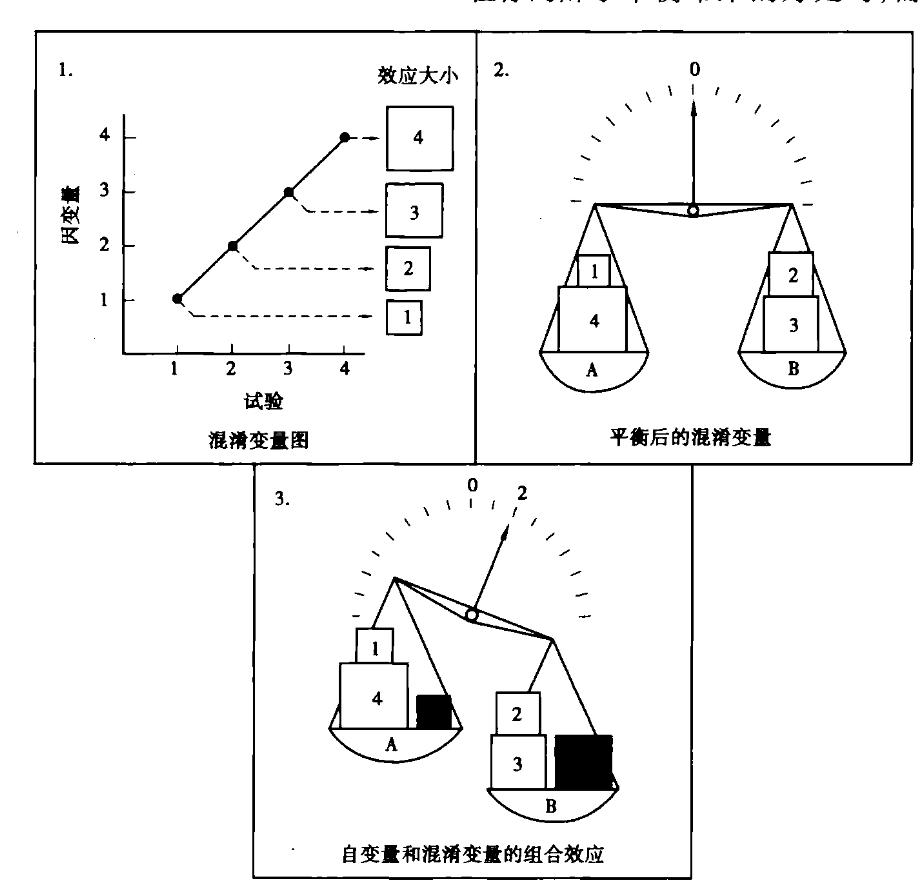


图 8-2 图左侧是线性混淆变量对因变量的效应。右侧是采用 ABBA 方法对混淆变量平衡之后的情况。下半部分是用阴影部分表示的自变量效应加入之后的情况，其中可以发现 2 个单位的净效应。

你的是，平衡方案的使用是基于某种前提假设的，如果违反了这些前提，使用平衡反而会起反作用。

ABBA 平衡法的一个前提假设是，额外变量对因变量的影响是线性的，即会形成一条直线。为了说明当不是线性时会产生怎样的影响，我们再回到重量的讨论中。假如额外变量的作用像图 8-3 的左侧(模块 1)显示的那样，事实上，学习效应是最常见的额外变量，许多学习曲线与模块 1 都很接近，在行为反应初期增长迅速，随后变化则相当缓慢。再转化为重量，根据 ABBA 的设计方式，将重量累积起来，在右侧(模块 2)中我们会看到天平是不平衡的，而是向 B 侧偏了 3 个单位，当再加上自变量的重量时，最后得到的净值是 5 个单位，而不是 2 个单位。

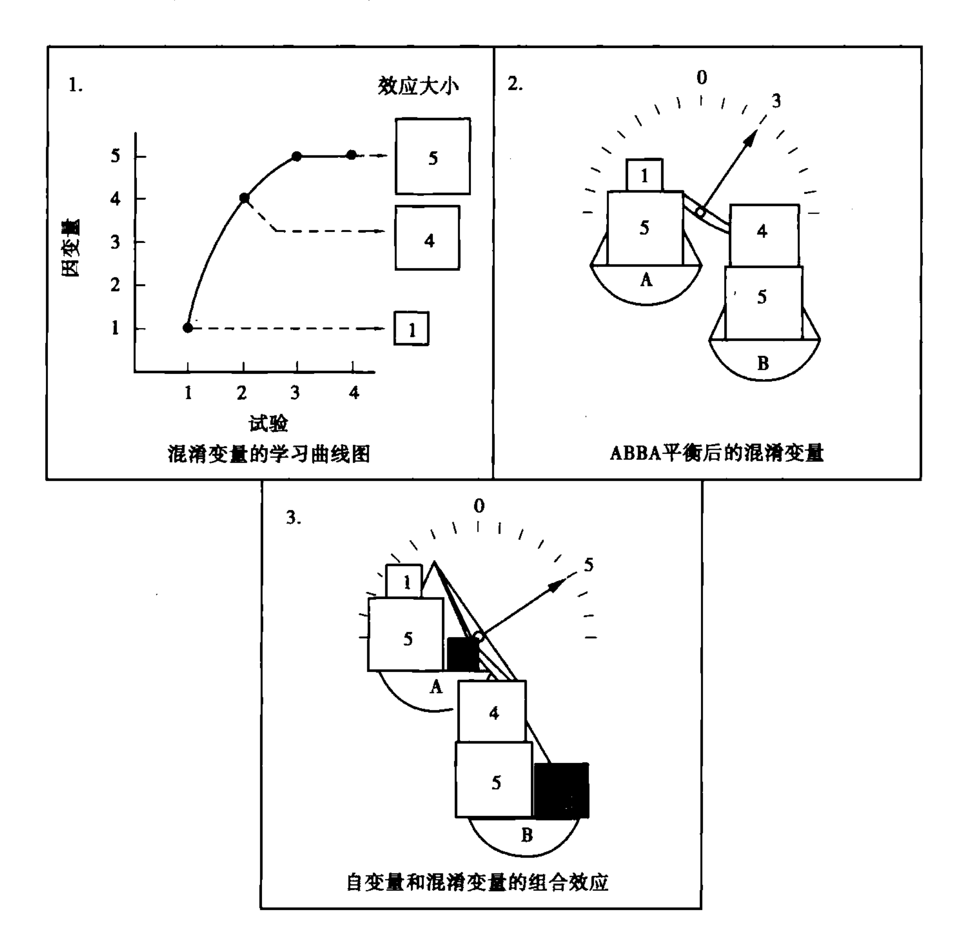


图 8-3　图中的模块 1 是学习曲线混淆变量对因变量的效应。模块 2 中，混淆变量没有被 ABBA 平衡掉，天平向 B 侧偏移了 3 个单位。下半部分(模块 3)是用阴影部分表示的自变量效应加入之后的情况，其中的 2 个单位的净效应被高估成了 5 个单位。

在某些条件下，ABBA 平衡法不仅不能矫正额外变量，有时还会使混淆因素对因变量的影响变得更加复杂。

这样的例子如图 8-4 所显示，混淆效应首先会提高行为反应，然后，降低行为反应，将学习效应和疲劳的作用组合在一起就产生了这样的曲线函数。接下来，我要让你计算出这种不平衡变量产生偏向的大小。

我们已经知道,只有当混淆变量的作用是线性时,ABBA 法才可以排除被试内设计实验中混淆变量的影响。如果混淆变量的作用不是线性的,我们必须选择一种不同的平衡方法或者采用被试间实验设计。

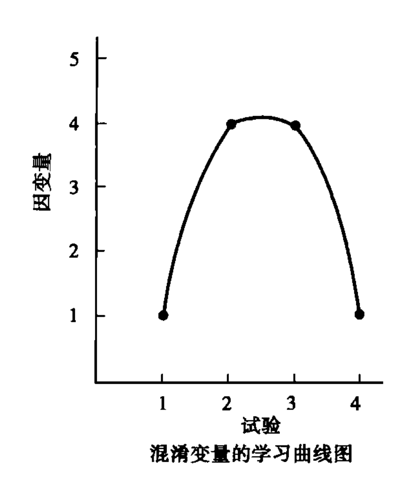


图 8-4 一个复杂的混淆变量影响因变量的示意图。这样的曲线可能是学习和疲劳引起。

ABBA 法是在完全被试内设计中平衡顺序效应的,被试既接受 AB 顺序,也接受 BA 顺序。其他平衡顺序效应的方法是将顺序作为一个被试间的变量,在最简单的单因素 2 水平实验中,一组被试接受 AB 处理,另外一组接受 BA,计算两组被试的 A 处理结果的平均值,也同样计算处理 B 结果的平均值。使用这种方法时,不一定要求混淆效应必须是线性的,但仍需要事先假设 B 在 A 之后产生的作用与 A 在 B 之后的作用刚好相反(Poulton & Freeman,1966)。这个假设有时被称为对称转换假设①。当违反了此假设,而得到非对称的转换时,这种平衡方式也是无效的。

仔细分析下面这个产生了非对称转换的实验。研究者感兴趣的是噪音对行为反应的作用有多大(Aldridge,1978; Poulton, 1979)。实验者先让被试识记一个辅音—元音—辅音组合(即 CVC,如 DOF)的字符串 16 秒;在完成记忆任务的同时,被试还会听到一系列的"B","B"每秒呈现一次,让被试去检测偶然出现的"P"。噪音条件下会出现一个音量较大的、连续的嘶嘶声。为了平衡顺序效应,一组被试先完成安静条件的试验,然后是噪音试验(AB);另外一组被试顺序相反(BA)。

图 8-5 是实验结果,先进行安静条件的实验组的记忆非常好。然而,当进入噪音条件后,被试的记忆成绩显著降低。另一个实验组的记忆成绩与预期一致,他们在噪音条件下的记忆成绩比较差。但是值得注意的是,当再将他们转入安静条件时,他们的记忆效果提高的幅度很小。先在安静条件记忆的实验组成绩降低了 31%,而先完成噪音条件记忆的实验组成绩仅提高了 10%。如果处理顺序的变化是对称转换的,变化量应该相等。出现非对称转换的原因是什么呢?

显然,这两个实验组的被试完成实验任务的方式不同。先完成安静条件的实验组可能使用回声存储器存储词汇,当声音刺激消失后②,回声存储器就像在大脑中的回声一样,短时间里会自动回响。如果马上出现一个响亮的声音刺激,它会将回声掩蔽。虽然回声策略在安静条件下效果很好,但当出现噪音后,先完成安静条件的实验组必须采用新的策略,如使用发音存储器(articulatory store)。这种条件下,他们要主动复述三字母,至少默读③。当转换为这种策略后,记忆效果大大下降。先完成噪音条件的实验组显然开始就使用了发音策略,当转入到安静条件后,他们

① 有时被称为并非不同的转换。

② 举个回声存储器的例子。当丈夫在看报,妻子问他"你是否听到我在说什么"时,他会使用回声存储器在记忆中搜寻。

③ 实质上是默读。

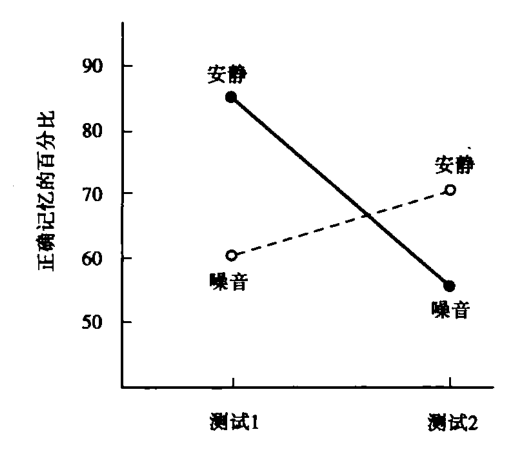


图 8-5 字符串记忆效果图。转换效应不对称。改编自："levels of processing in speech perception", by J. W. Aldridge, 1978, Experiment 4, Journal of Experimental Psychology: Human Perception and Performance, 4, 164-177.

仍使用这种效果并不好的策略，因此，他们在没有噪音的情况下，成绩改善并不大。虽然这样的解释只是一种推测，但我不想再拿更多支持该推测的证据或数据来使你感到厌烦。

还有一个例子可以帮助你理解非对称转换。假设你想考察饮酒对复杂运动功能的影响，如玩视觉游戏中的赛车。你让一组被试喝完 3 杯酒，去开 3 次赛车，每次 1 个小时。然后，转入清醒状态下再开 3 次。为了控制顺序效应，让另外一组被试按相反的顺序完成任务，先在清醒状态下开 3 次，饮酒后再开 3 次。你预期每组被试在饮酒状态下得分较低。你会惊讶地发现，先饮酒组在转入清醒状态后最初在行为上是下降的，但不会像另一组酒精的作用那么大。这个结果就是学习依赖状态的非对称转换的例子。我们在某个特定的状态下（如清醒或醉酒）学习某个技能，当我们再回到当初的状态上时，我们倾向表现最好。或许你知道某人在喝酒后射击会更准，这就是依赖状态的学习。在我们的实验中，依赖状态的学习会导致非对称转换，就像我前面所讨论的那样。当有非对称转换发生时，没有任何一种平衡方法可以补救被试内设计。

当自变量的水平增多时，完全平衡法的复杂度也增加。在一个完全平衡设计中，每个水平都要发生相同次数。表 8-4 显示了两个水平、三个水平和四个水平实验的完全平衡设计。当单个自变量有多个水平或者多个自变量时，完全平衡会变得非常庞大。在大型实验设计上，要像第 2 章所描述的那样，随机地分配水平或使区组随机。有时，还可以使用部分平衡法，此时，你只需要平衡某几个顺序，以确保每个水平在每个位置上发生相同的次数。

一种常用于两个水平以上的自变量的平衡方法是拉丁方设计（Latin Square），拉丁方设计保证每个水平出现在不同顺序位置上的次数相同。对于自变量水平数量一定的拉丁方设计，可以有多种不同形式，更常用的是平衡的拉丁方设计，在这种设计中，不仅每个水平出现所有顺序位置的次数相同，每个条件处于其他条件之前

表 8-4　两水平、三水平和四水平的完全平衡设计

两个水平自变量		三个水平自变量	
数　目	水平顺序	数　目	水平顺序
1	AB *	1	ABC
2	BA	2	ACB
		3	BCA
		4	BAC
		5	CAB
		6	CBA

四个水平自变量			
数　目	水平顺序	数　目	水平顺序
1	ABCD	13	CABD
2	ABDC	14	CADB
3	ACBD	15	CBAD
4	ACDB	16	CBDA
5	ADCB	17	CDAB
6	ADBC	18	CDBA
7	BACD	19	DABC
8	BADC	20	DACB
9	BCAD	21	DBAC
10	BCDA	22	DBCA
11	BDAC	23	DCAB
12	BDCA	24	DCBA

* 字母 A,B,C 和 D 分别代表水平。

和之后的次数也相同。假设我们想知道阅读计算机屏幕上 4 种字体(Chicago,Courier,Geneva,Times)的文本各需要多长时间？考虑到呈现的顺序会使实验变得复杂，图 8-6 表示了这个实验的平衡的拉丁方设计方法，请注意 4 个被试阅读 4 种字体的顺序。以 Courier 字体为例，不同行中位于它前面的字体分别是 Chicago，无，Geneva 和 Times，位于它后面的字体分别是 Geneva，Times，Chicago 和无，这些都满足平衡拉丁方设计的要求。在实施拉丁方设计实验中，还要求被试至少与自变量的水平数相等，简单而言，应该是自变量水平数的整数倍。这种形式的部分拉丁方设计考虑到了绝大多数由非对称转换产生的混淆因素，但也带来顺序和非对称转换之间的交互作用而导致的混淆效应。全平衡设计是最好的，但当被试非常少的时候，使用拉丁方设计更好。

	呈现顺序 第 1	第 2	第 3	第 4
被试 1	**Chicago**	Courier	Geneva	Times
被试 2	Courier	Times	**Chicago**	Geneva
被试 3	Times	Geneva	Courier	**Chicago**
被试 4	Geneva	**Chicago**	Times	Courier

图 8-6　4 名被试在 4 种字体呈现顺序的平衡拉丁方。

你已看到，平衡的方法可以用于最大程度降低被试内设计产生的顺序效应。但还应该注意使用这种方法的前提假设，只有在满足这些前提假设的情况下，才能使用平衡法。然而，在某些如含非对称转换实验中，那些假设是不可能满足的，此时你别无选择，只好使用被试间设计了。如果被试内设计的缺点不能通过平衡法纠正，那么，你可能必须使用另外一种被试间设计——范围效应。

范围效应

假设你是一名小机械厂的采购员，你想为机械生产线购买一种新型的工作台，你必须仔细挑选工作台的高度，并且确保工作台的高度是最适合高产量的。你想做一个实验来确定工作台的高度，你找了一组工人，即组 A，让他们坐在高度不同的工作台旁，计算他们在 3 分钟之内能够把多少块积木翻过来。选择的台面高度分别比肘部高 -10，-6，-2，+2，+6，+10 英寸。看过本书以后，你知道可能会遇到顺序效应问题，所以，你小心翼翼地平衡不同高度的顺序。

实验完成后，你的老板表示，她想让你再测试一下高度更低的工作台，所以，像刚才的实验那样，你又设计了一个新的实验，不同的是，这次你使用了另外一组工人，即组 B，台面的高度变成了 -18，-14，-10，-6，-2 英寸。

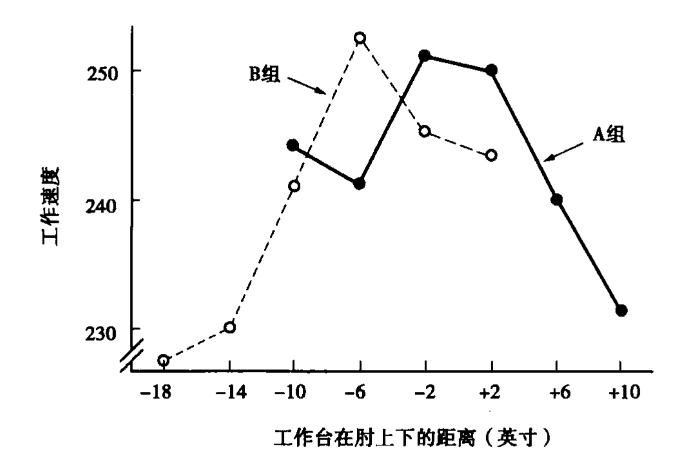


图 8-7 3 分钟内工作台不同高度条件下的距离效应。图中表明，当工作台高度在中等情况下，工作效率最高。引自："Series Effects in Motor Performance Studies," by J. E. Kennedy and J. Landesman, 1963, *Journal of Applied Psychology*, 47, 202-205.

实验结果见图 8-7。令人感到奇怪的是，两组被试最适宜的高度是不同的，A 组被试最适宜的高度大约就是肘的高度，B 组被试则是肘部以下 6 英寸的位置。为什么会这样呢？因为被试学会了在特定高度的工作台上，完成翻积木的任务需要的技能，同时还学会了在其他高度中也需要的技能。两个高度越相似，被试越容易将这种技能从一种高度迁移到另一高度上。这正是学习的一个基本原则。因此，我们仔细分析一下，这个积木翻转实验也是一个学习实验，我们希望工人能够在每个高度上都尽其所能地完成任务。表 8-5 显示了不同高度上的差异。例如，A 组在 -10 和

+10 高度差是20 英寸，+10 和 -6 的高度差是16 英寸等。把A组其余5 种高度与+10 的高度差加在一起，除以5 后得到均值为12。如果我们希望两个实验中的最高效率相同，我们可以从表8-5 中预测出图8-7 来。此时，你可以看出为什么称之为“范围效应”了。被试在自变量水平范围的中间位置上完成得最好，因为在水平范围的中间部位上，学习迁移的程度最高。范围效应产生于被试内设计的实验中，而不论此时刺激和反应是否每次都按一致的顺序呈现。Poulton(1973)发现，实验心理学的绝大多数领域中都存在范围效应。

表8-5 每种工作台高度与其他5 种工作台高度差的平均值(英寸)

	工作台高度							
	-18	-14	-10	-6	-2	+2	+6	+10
A组			12	8.5	7.2	7.2	8.5	12
B组	12	8.5	7.2	7.2	8.5	12		

虽然 Poulton 和一些研究者都警告，慎用被试内设计，但其他研究者则认为，许多情况下，都应该首选被试内设计。如 Greenwald(1976)指出，范围效应仅是一种上下文效应(context effect)，被试来参加实验时，早就已经形成了上下文情景。如工作台的例子中，被试已经使用过某种高度的工作台。他认为，像在被试间设计的实验那样，反复给被试呈现自变量的某个水平，并不能消除上下文效应。重复自变量的一个水平，会产生新的上下文——即只有一种水平的上下文情景。由于这些原因，Greenwald 认为，不论使用哪种设计，上下文效应都是不可避免的。他还提出，在选择设计类型上，更重要的问题应该是将实验结果推广到何种情景中去。

例如，在暴力电视的实验中，与让儿童接受多种水平节目的处理相比，只让儿童重复接受暴力节目水平的处理(被试间设计)更加不符合实际情况。我们希望将实验结果推广到具有多种处理水平的现实生活中，因此，我们更应该选择被试内设计，这时在实验中所使用的自变量范围与我们想要做推论的情景更相似。那么，作为一个实验者，尽管需要注意范围效应对实验结果的改变，也需要选择可以更好地将实验结果推论到合适情景中的设计类型。

Poulton 认为，被试间设计更好；而 Greenwald 以及其他研究者则认为被试内设计更好。这两种设计方式到底哪个更好呢？一个较为合理的观点是，这要依赖于你要进行怎样的实验。有些情况下，如在态度形成和某些记忆领域以及其他一些研究领域中，即使是周密的平衡方案也无法矫正如非对称转换之类的顺序效应，因此，就不可能使用被试内设计。在其他情况下，如已经发现某种心理治疗技术非常有效，此时也不能使用被试内设计。因为被试内设计可能会使它的疗效反转。而另一方面，使用一个被试作为他自己的控制条件是一种非常好的实验手段，它可以降低我们认为微不足道但对实验控制又非常重要的变异。有些领域的研究，如在记忆研究中考察保持时间和记忆负荷，在注意研究中考察启动，在知觉研究中考察表象和声音定位等，在这些研究中使用被试内设计，即使存在顺序效应问题，但其影响并不大。这些情况下，更有效的设计方式应该是被试内设计。因此，作为一个研究者，你要做的最正确的事情是选择一种最合适你研究问题的实验设计。

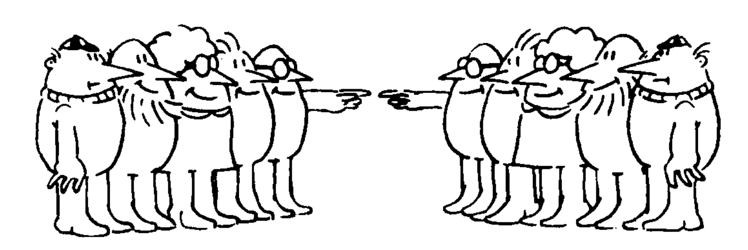
匹　配

如果既想利用被试间设计的优点,又想避免不同实验组之间被试的个体差异问题的方法之一是使用匹配组。简单而言,就是试着为自变量的每个水平都分配相同类型的被试。在典型的被试间设计实验中,你希望分配到不同水平上的被试非常相似,所以,采用随机化的办法,随机分配可以将实验组变得更加同质,尤其当实验组规模非常大时更是如此。然而,由于采用了随机分配的方式,偶尔也会出现被试分配有非常大的差异,这时你会错误地将这个差异加到了自变量上。你的实验就出现了实验组的差异这一混淆因素。通过使用匹配组,可以使这种情况发生的可能性降到最小。

那么,我们依据什么因素进行组间匹配呢?你必须依据与因变量高度相关的变量来匹配实验组。在跑鞋的实验中,如果根据 IQ 分数来匹配实验组,那肯定是在浪费时间,思维敏捷与步履敏捷没有直接关系。我们可以先让每个被试穿网球鞋跑一下;然后,按照下面方法匹配被试,即第一名和第二名是一对,以此类推。我们可以通过掷硬币确定将每一对中的哪名成员分配到哪一个跑鞋组。通过这种方法,我们可以知道,在引入自变量前,两个组在奔跑速度上大致相同。在这个实验中,我们的假设是,穿网球鞋的速度和穿短跑鞋的速度非常相关。如果匹配变量与因变量相关程度非常低,那么,进行匹配的意义就不大了。

匹配组设计

如果使用了匹配,当我们说不同的自变量水平产生了行为反应上的差别时,犯错误的可能性就会大幅度降低。匹配还带来统计上的好处,当使用了匹配组,统计检验的结果得出的因变量差异更有可能不是由随机误差造成,而是由自变量产生的,即统计检验对自变量引起的任何差异变得更加灵敏。



匹配组设计

为了详细说明这个规则,表 8-6 的左栏中列出了从表 8-2 中随机选择一组被试,让他们穿钉长为 7 mm 的钉鞋。假设我们已经让他们穿网球鞋短跑过,括号中的数字是网球鞋的成绩。为了得到匹配的实验组,我们现在让更多的人穿网球鞋短跑,我们选那些与最初选中的被试速度相同的人,新的被试列在右栏,括号中也是他们的网球鞋成绩。请注意,我们降低了两组被试穿网球鞋跑步速度的差异,使他们完全相同。接下来,我们让新的一组穿钉长为 13 mm 的钉鞋短跑。结果表明,与我们

前面的实验结果相同,即平均时间存在0.5秒的差异。你是否更加相信是鞋钉的长度造成了两组奔跑速度之间存在0.5秒的差异呢?通过统计检验,也会得到相同的结论①。

表8-6 两个匹配组的百米短跑时间

穿7 mm钉长跑鞋的男子	时间(秒)		穿13 mm钉长跑鞋的男子	时间(秒)	
Mike	(12.2)	11.7	Vic	(12.2)	12.2
Homer	(12.8)	12.2	Jack	(12.8)	12.6
Randy	(13.5)	12.9	Barry	(13.5)	13.5
Gordon	(14.0)	13.2	Larry	(14.0)	13.8
John	(14.3)	13.7	Jess	(14.3)	14.2
George	(16.1)	15.4	Stuart	(16.1)	15.8
Harry	(16.7)	15.8	Harvey	(16.7)	16.2
Tim	(17.0)	16.0	Sid	(17.0)	16.6
Bob	(18.7)	18.2	Pat	(18.7)	18.7
Bill	(19.7)	19.1	Joe	(19.7)	19.6

穿7 mm钉长跑鞋的男子的平均速度=14.82秒;穿13 mm钉长跑鞋的男子的平均速度=15.32秒。平均差=0.5秒。

使用匹配组方法的一个缺点是,匹配实验组要耗费很长时间,因此,造成实验通常被分成两个部分,一部分是前测,另一部分才是正式实验。如果你计划用大量的被试,使用随机分配的方式造成实验组之间有较大差异的可能性非常小,因此,不一定值得花很大力气去进行匹配。

最后一个要考虑的问题是匹配过程本身也可能会产生一些问题。在上述例子中,我们假设穿网球鞋的前测对钉鞋测验不会产生很大影响。假如网球鞋测试教会了被试穿平底鞋的奔跑技巧,这个技巧会迁移到后面的测验中。因此,可以预期钉鞋越平,跑的速度越快。因为短钉鞋与平底鞋更相似,因而速度更快。这种情况下,前测会造成了被试在后测自变量的两个水平上的不同。

因此,在某些条件下,匹配组设计是有价值的,但也会产生很多问题,而不是解决问题。在实验中,你要权衡一下使用匹配设计的利与弊。表8-7总结了本章讨论过的各种设计的优点和缺点。很明显,在做任何实验时都要认真考虑这些利弊。当我们在第9章中讨论多变量实验时,你会发现,在很多情况下,一个实验中可以同时包含被试间设计的变量和被试内设计的变量。例如,如果我们对"提示灯会对声音的反应时产生的作用"是否受到被试性别的影响感兴趣,我们需要将"有提示灯和无提示灯"作为被试内变量,而性别作为被试间变量②。

① 注意:使用匹配组设计进行统计检验时,已经假设你成功地在与因变量高度相关的变量上做了匹配。正因为如此,统计检验结果上显示存在显著差异的结论也更加稳妥。如果你匹配的是与因变量相关不高的变量,那么,你得到显著差异的可能性并不比当初根本没有匹配时更高。

② 到底将变量设计为被试内还是被试间,这个问题对于有些变量毫无意义,如性别、种族、人格特质以及IQ分数等。

表 8-7 被试间设计与被试内设计优缺点的总结

组内实验设计	
优 点	缺 点
需要被试数量少。	条件之间可能出现转换效应。
实验时间短。	ABBA 平衡假设混淆效应是线性的。
组间变异较小。	所有平衡法都假设转换效应是对称的。
	范围效应导致新问题。
组间实验设计	
优 点	缺 点
条件之间不可能出现转换效应。	可能存在组间差异。
不需要平衡。	需要更多被试。
匹配可以降低组间变异。	实验时间长。
随机分配被试可以消除偏差。	匹配费时费力，并假设其中不存在转换。

因此，对于多变量实验而言，"被试内"和"被试间"指的是变量，而不是实验。

小 结

将被试分配到自变量的不同水平上有两种基本方法，即可以将不同被试分配到每个水平上，或将相同的被试分配到所有水平上。第一种方法是被试间设计，第二种是被试内设计。被试间设计的一个优点是，被试只接受自变量一种水平的处理，因此，其他水平的处理不会影响被试的反应；此外，实验时间会比较短。被试内设计的最大好处是由被试个体差异造成的变异得到了最大幅度的减少。在实践中的其他一些优点是，被试内设计需要的被试量较少，因此，最大程度地减少了对被试的训练时间。被试内设计的缺点是，需要平衡其中的顺序效应。ABBA 平衡法可以平衡被试自身的顺序效应，但前提是必须假设顺序效应对因变量的作用是线性的。也可以使用完全平衡法，但必须假设不同条件是对称转换的。在较为庞大的实验中，完全平衡是不可能的，此时可以使用部分平衡法、随机分配或区组之间进行随机等。在刺激或反应总是存在一定的顺序效应时，即使使用平衡法也不会克服实验中的范围效应。在将被试分配到不同实验组接受某一水平的处理时，匹配组的方法可以降低被试的个体差异。

9

如何计划做一个单变量、多变量和聚合序列实验

如果实验所依据的科学假设毫无价值，那么，无论多么仔细的实验构思和实施过程都没有任何意义。

R. E. KIRK(1968)

我仍必须关注可能存在的任何问题，无论多么复杂，当你用正确的方式去认识它时，它才不会变得更复杂。

PAUL ANDERSON

本章我们首先讨论单变量实验，到此为止，这种实验设计构成了本书中几乎所有的例子，这些实验都只有一个被操纵的变量，它们要么有两个水平，要么有多个水平。然后，讨论多变量或多因素的实验。在多因素实验中需要同时操控几个变量，每个变量又分为两个或多个水平。在心理学著作中，多因素设计的实验比其他类型的实验更为常用。最后，我们将讨论聚合序列设计，在聚合序列设计中，包含多个单变量或多变量的实验，以验证某个假设或理论。

单变量实验

两水平的实验

最简单的实验只有一个自变量，且该自变量只有两个水平。有些研究者将这两种处理水平分别称为实验组和控制组。在有些情况下，显然，控制组应该是没有任何实验处理的。例如，假设你对"某种药物对行为的影响"感兴趣。那么，实验组服药，控制组则不应该服药。控制组在这种情况下仍然非常有用，它能够显示实验条件是否是产生效应的原因。在其他情况，尤其是当自变量有多个水平时，到底将哪个水平称为控制水平①不是很清楚，正因为如此，我仍然坚持使用"水平"来描述自变量。不论在何种情况下，一个真正的实验至少必须有两个水平。否则，因为没有比较，就不可能说自变量的变化引起了行为反应的改变。

在实验心理学的早期历史中，报告的典型实验都是两水平的单个自变量实验。因为在那时这门学科还很年轻，研究者们对发现某个自变量是否有作用更关心，而不是确定这种作用的本质规律。另外，那时也缺少能够分析更复杂实验设计的统计方法。尽管存在一些统计方法，但一般研究者也不是很熟悉。

现如今，杂志编辑们通常希望在一个实验中看到两个以上水平的设计。有时，他们也认可做得非常好的两水平设计，尤其是包含几个连续的实验，但现在典型的实验都包含多个水平。不过，作为第一个实验，设计两水平的实验比较适合。新手首先需要湿一下鞋子而不被淹死，有些情况，两水平实验也可以提供有价值的结论。

优　点

实际上，两水平的实验确实具有复杂设计的实验所不具备的优点。它提供了确定某个自变量是否值得研究的方法。如果某个自变量对人的行为反应没有任何作用，显然，就没有必要再设计一个更复杂的实验来考察这种作用。

两水平实验的结果也易于解释和分析，结果仅仅是，"是的，这个自变量确实有作用；行为反应是在这种方向上变化的"；或者"不，这个自变量没有作用"。通常需要通过统计检验确定变量的作用是真实存在，还是由随机误差引起。对于两水平的实验而言，这种检验的方法非常简单，并不比加减法运算复杂多少。一旦你知道应该使用哪种统计方法，只需要花几分钟时间操作一下计算机，就可以分析出数据

① 例如，实验中我们以性别作为自变量，到底是将男性还是女性作为控制组呢？男性主义者和女性主义者对此争议了许久，所以何不避开这样的争议，而分别称为水平 1 和水平 2 呢？

结果。

最后，有些时候两水平的实验可以为你提供足够信息。如果实验目的是想验证两种冲突的理论观点，一种理论预计在两个水平上存在行为差异，而另一个则预计没有差异或差异的方向相反，那么，两水平的实验就足以分辨出两个理论的对错了。此外，在有些应用研究中，两水平实验也可以提供非常有价值的信息。例如，只有两个水平比较重要或只存在两种水平时，你可以考察两种心理治疗技术、两种教育方法、两种训练步骤、两种药品、两种性别或其他任何变量的两个水平的差异和作用。

缺　点

虽然两点之间直线的距离最短，但直线并不是两点之间存在的唯一线。也就是说，许多两水平实验也存在缺陷，因为它们不能得出自变量和因变量之间关系的具体模式。

假设我们想做一个实验，看一下本书应该使用多大的字体才能使读者尽可能快地读懂我这冗长的表述。我们会使用几个段落，一些段落的字体是 12 号，其他的用 10 号。然后，我们可以测量人们阅读不同字体段落的时间。当然，我们需要控制段落的难度和顺序效应，需要使用本书介绍的其他各种好方法。

实验结果见图 9-1。两个数据点之间的任意直线表示，字体越小，阅读时间越长。因此，实验已经回答了我们的问题：用 12 号的字体可以使读者阅读更快。然而，如果我们真正想知道，在众多不同大小的字体中，哪种字体是最佳的，那么，上面的实验就不能使我们得出这样的结论了。我们的结果并不等于说，“字体大小和阅读时间之间的关系是直线关系”，这样的结论也并不一定对所有字体都是正确的。

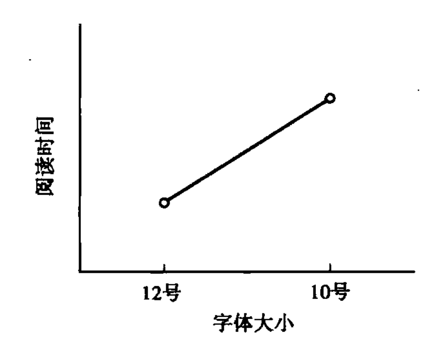


图 9-1　阅读一段字体分别为 10 号和 12 号的文字的时间

图 9-2 显示了可能存在的其他关系类型，你会发现，在不知道二者关系的形式时，会导致结果的内推存有问题①。

从两点之间进行外推比内推将更加危险。许多心理学函数曲线都有“天花板效应”和“地板效应”。当因变量达到一个不可逾越的水平时，就会产生“天花板效应”。典型的天花板水平是反应的正确率为 100%、发生的可能性为 1、产生某个反应的确信度为 100% 等。在这些情况中，没有人会产生超过上述值的反应(换句话说，你不会比 100% 更准确的了)。而在其他情况下，虽然极值没有限制反应，但有

---

① 内推是在我们现有的知识范围内进行直接估计，外推则是在现有的知识范围外进行估计。

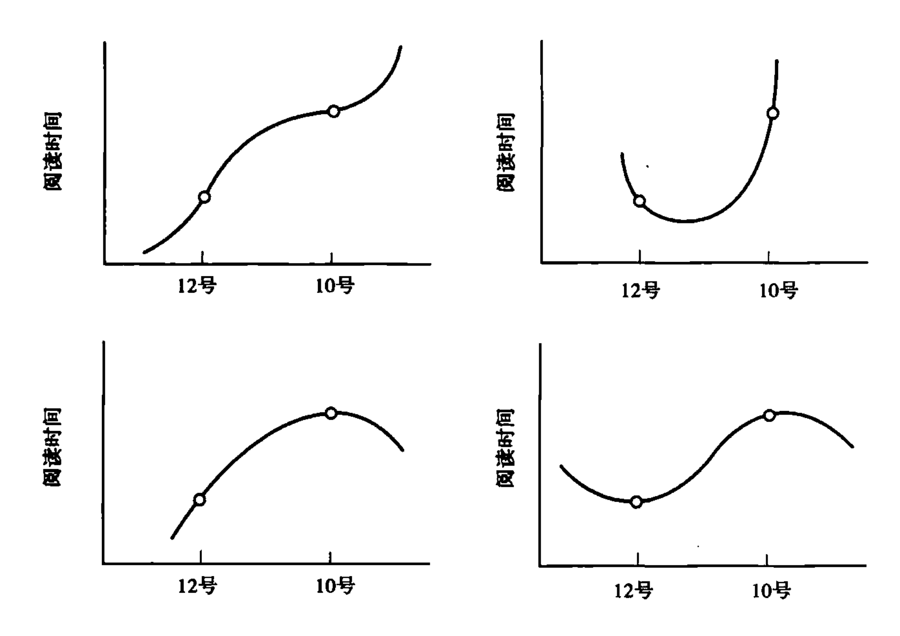


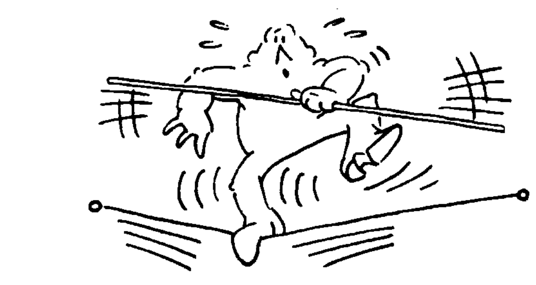
图 9-2　阅读时间与字体大小之间的几种可能的关系。所有关系的曲线都经过相同的两个数据点

些天花板效应会限制反应，如即使大量练习，我们在短时记忆中保持的项目数量就有一个限制——大约是 7 个项目。同样，在有限时间内，人们可以有效加工信息的数量也是有限的，如求和时的数字个数、在一次呈现中可以识别的目标个数、可以打印的单词数等。“天花板”并非牢不可摧，但它始终难以逾越。

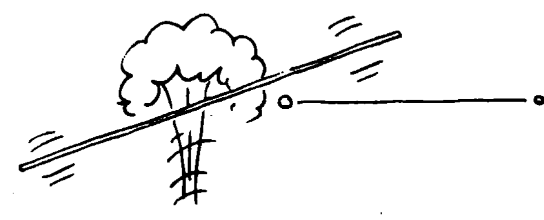


“地板效应”表示不会产生低于某个值的反应。例如，人的反应速度不可能低于 0 秒，或产生比不作任何反应还少的反应。同样，地板也并非是绝对的，也有一定的弹性。例如，虽然从理论上来讲，觉察某个刺激的最短时间是零，但最低限度大概可能是 150 ms。如果我们把两个数据点之间的结果外推到高于天花板或低于地板水平，那么，结论是否正确就可想而知了。有时，天花板或地板效应可能不太明显，但为了避免这些问题，在两水平实验中，“不要将结果在实验水平范围之外进行内推或外推”，应把它作为一个原则来遵守。

有时，两水平实验在理论上存在局限。我们同意第 1—3 章所描述的那样，科学是建立在某种关系之上的，可以用某种理论去解释实验的发现。每个理论都会与其他可能的理论存在相互矛盾的地方。但通过实验，结果支持其中之一，也意味着排除了其他理论。因为许多理论会预测，自变量的变化导致因变量在某个方向上的变化，而两水平实验的结果通常不能分辨相矛盾的理论。除非相矛盾的理论预测的方向相反，或一个预测有差别而另一个预测没差别。因此，理论的检验通常需要更复杂的实验设计。



两点之间的内推很冒险!



两点之外的外推更危险!

多水平实验

多水平实验是指多于两水平(三个或更多)的单自变量实验。许多研究者也称之为"函数实验",因为这些实验可以得出反映自变量和因变量关系的函数曲线。

优　点

多水平实验的主要优点是,实验结果可以让我们推断自变量和因变量关系的本质特征。即使一个实验只有三种水平,与两水平相比,它仍能够更好地说明自变量和因变量的潜在关系。

假设我们想知道学生的焦虑水平如何影响成绩,我们决定使用两种类别①或两水平的被试间设计。第一类是在学生考试前,教师用 5 分钟时间向学生讲述成绩对于学业的重要性,让学生清楚地知道,最好的成绩会得到最好的工作,获得奖学金的同学未来会有非常高的薪水,而且当前考入大学的竞争很激烈。

第二类是学生考试前,老师也给予 5 分钟的谈话。但这次是告诉学生,此次考试成绩就像平时学习知识一样不甚重要。并告诉他们,10 年后,他们将不会再记起这次考试。在这个实验中,我们尽量仔细控制各种混淆变量,如成绩水平、测验的难度以及每个类别的指导语等。因此,我们认为,测验成绩的差异可以归因于谈话引起的焦虑水平。假设第一种谈话引起了高焦虑,第二种是低焦虑,我们会得到图 9-3 的结果:

到此为止,我们能做的最好推测是焦虑水平和成绩之间没有关系,高焦虑和低焦虑两点之间的直线是水平的。假设我们决定使用多水平的实验设计,加入了第三种焦虑水平,一种自然条件,老师用 5 分钟的时间简单地提醒学生一些考试细节,图 9-4 显示了此多水平实验的假想结果。

---

① 因为这里我们使用已经存在的两种类别,而不是将学生进行任意分类,这个例子不是一个实验设计,而是一个准实验设计(将会在第 10 章中讨论)。希望读者注意此差别。

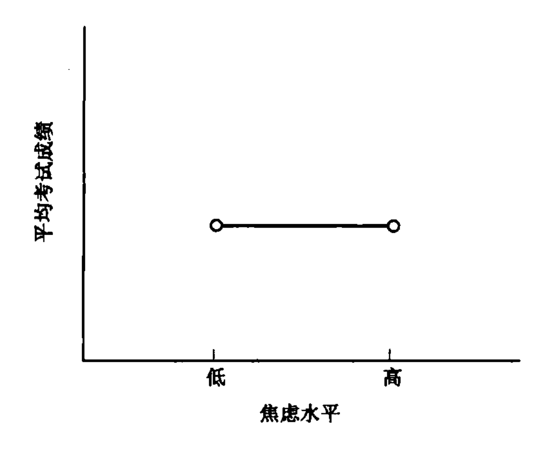


图 9-3　学生焦虑水平和平均考试成绩实验的假设结果

当我们画出了第三个数据点后发现，虽然对这种关系的真实形式还存有疑问，焦虑水平和测验结果之间的确存在重要关系①。我们知道心理学曲线不是骤然变化的，图 9-4 中 3 种形式中的任何一种都有很大可能会发生。因此，第三个数据点使我们更好地考察自变量和因变量关系，当我们陆续加入更多的水平后，我会对二者关系作出更好的推测。我们更有信心地从数据中进行内推和外推。在这个实验中，我们加入的第三组可以看成是控制组，因为此时教师没有试图去影响学生的焦虑程度。还可以加入另外一种控制组，即教师什么也不说，那么，这个控制组可以用来确定考前谈话是否会影响成绩。因此，多水平实验具有一定的灵活性。

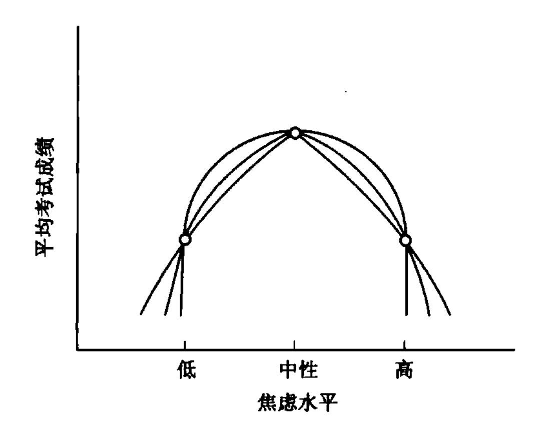


图 9-4　学生 3 种焦虑水平和平均考试成绩实验的假设结果

这个例子也说明了多水平实验的第二个优点，即一般而言，水平增加得越多，自变量的范围也就越不那么至关重要。在第 7 章中曾经讨论过，虽然自变量的范围选取要切合实际，但如果要考察可能存在某种关系时，也可以选取更大的范围。显然，当实验具有多种水平时，这些要求都很容易得到满足。

---

① 如果你具有动机或注意方面的知识，你会注意到，这是 Yerkes-Dodson 定律，即唤醒程度和学习效率之间是倒 U 型关系。

缺点

从实践角度来看,多水平实验的主要缺点是,它比两水平实验需要更多的时间和精力。当我们每次增加被试间实验水平时,我们都需要增加被试数量。在被试内实验中,水平的增加虽不会导致被试数量的增加,但却增加了实验的总时间,它使平衡方案变得更烦琐。

多水平实验数据的统计分析也非常困难,需要更多的时间,结果也更难从统计的角度去解释。

权衡两水平和多水平实验的优缺点,增加自变量水平所需要的投入较小,通常可以获取更有价值的信息。这种益处对开始增加的几个水平尤为明显。当然,在某个转折点处,增加更多的水平也并不会增加我们对自变量和因变量关系的认识。

到此为止,我们假设所有的实验只有一个自变量,然而,这更多的是一种局限,而并非完全是对现实世界的反映。许多实验会含有多个自变量,接下来,我们将要讨论在设计更为复杂的实验过程中常用的一般策略。

在实验心理学中最常用的实验设计是因素设计(factorial design,亦称因子设计或析因设计——译者),为了能够理解心理学杂志中绝大多数实验的结果,你必须理解因素设计的逻辑。

因素设计

将多个自变量放在一起的典型方法是因素组合,即将一个自变量的每种水平与第二个或第三个等变量的各种水平进行匹配,在这种设计中自变量被称为因素①。

请看一个因素实验(factorial experiment)的例子。假设你想知道有领袖的群体是否要比没有领袖的群体达成一致意见的速度更快?你需要考虑控制哪些情景,改变哪些因素。例如,群体成员是否应该是同一性别?沟通是结构式的还是自由的?该让群体去处理一个容易的还是难的问题?你会发现,很难对上述因素进行严格控制或随机处理。例如,你会觉得领袖对群体的领导效力取决于群体的规模,因此,你要选择是否有领袖和群体规模作为因素。假设这两个因素的水平分别为领袖(有、无)、群体规模(3 人、6 人、10 人和 20 人)。

图 9-5 是因素实验常用的表示方法,矩阵的每个边由一种因素组成。矩阵中的每个方格称为"单元",与更为简单的实验类似,被试被随机分配到不同单元中。在该例中,左上角的单元中分配 3 名被试,其中 1 人是领袖。每一行或每一列都可以形成一个单自变量实验,我们称上面这个例子为 2 × 4 设计②,因为其中一个因素有 2 种水平,另一个有 4 种水平。

在因素设计中,只有你的想象力和世上的人数才是因素的数量极限。假设我们

---

① 有的研究者也称之为处理,这样会产生处理组合。在构建这个学科时,我们也存在像"巴比伦塔"这样的圣经故事,没有语言上的完全统一。新的措辞或许会使新手感到非常茫然。

② 表述中的 × 读做"by",而不是"times"。因此,对这个设计的英语读法(而不是阿拉伯读法)为"a two-by-four design"。

认为，群体决策时间不仅受有无领袖和群体规模的影响，还随群体成员的性别而变化，那么，我们就可以将成员性别作为具有 3 个水平的第三个因素。是 3 个水平吗？对，男性、女性和混合（男女各半）。图 9-6 是新的实验设计①，我们称之为 2×3×4 的因素设计。

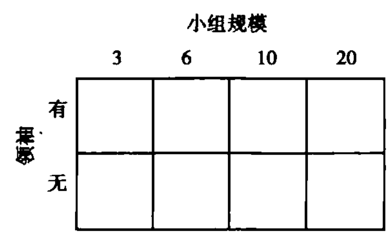


图 9-5　一个 2×4 的因素设计模式图。一个因素是领袖，包含有和无 2 个水平。另一个因素是小组规模，有 3,6,10 和 12 成员等 4 个水平。

在第 8 章中我们讨论了被试内实验和被试间实验，因为只有一个自变量，所以，这种称谓是可以的。但在因素实验中，因素本身也可以是被试内的或被试间的，或者两种类型同时包含在因素实验中，此时我们将实验称为“混合因素设计”实验。例如，在领袖实验中，我们可以将不同类型的群体分配到单元中，让每个因素都是被试间因素。我们也可以使用混合设计，让每个群体有或没有领袖，这种情况，群体规模是被试间因素，有无领袖是被试内因素。在决定将某个因素作为被试内还是被试间处理时，其利弊就像第 8 章讨论的那样，应该仔细考虑。如果需要，也必须对被试内因素进行适当平衡。

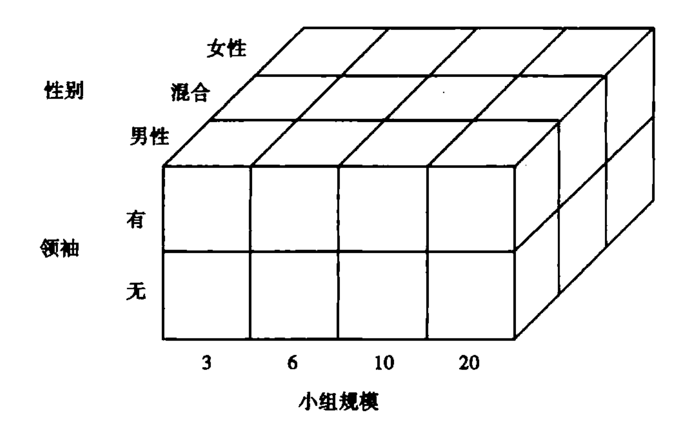


图 9-6　2×3×4 的因素设计模式图。因素有领袖（有、无）、性别（男、混合和女）以及小组规模（3,6,10 和 12 人）。

优　点

因素设计的主要优点是我们可以研究交互作用。当一个自变量和行为反应之间的关系依赖于第二个变量的水平变化时，就会出现交互作用。例如，3 人的群体不论有无领袖都比较容易地作出决策。但是，随着群体规模的增加，没有领袖的群体达成一致意见的时间更长。因此，有无领袖与决策时间的关系依赖群体规模的影响。图 9-7 显示了这种交互作用，你可以看到，对于 3 人构成的群体，是否有领袖不会影响解决问题的时间。当群体增大时，是否有领袖对减少决策时间变得越来越重

---

①　尽管形象地描述超过 3 个因素的实验非常困难，然而，实验设计类型不受三维空间的限制，只是比二维的难画些罢了。

要。两个单一自变量实验无法提供交互作用信息,而只能简单地使我们发现,领袖或群体规模是否有作用。只有在因素实验中,才能考察交互作用。

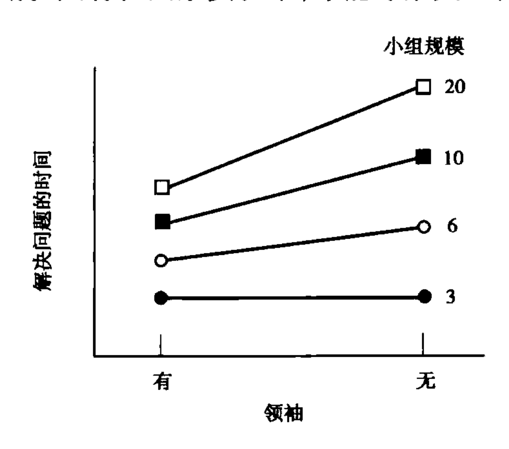


图 9-7 假设的结果显示,领袖和小组规模之间可能存在交互作用。注意:对规模最小的组而言,解决问题的时间与领袖无关。但对规模较大的组来说,领袖使解决问题的时间减少。

还记得在第 2 章中,我们讨论过有限的环境因素能决定行为吗? 通过实验,我们可以选择其中的一种因素作为自变量,其他环境因素要么被控制,要么按随机的方式变化。一旦我们确定了这种因素对行为的作用后,我们就可以选择其他因素进行研究。这种方式存在的问题是这种朴素的假设,即一旦我们知道了每个自变量的作用,就可以把这些因素相加在一起去解释行为。这样的假设完全忽视了环境因素之间的交互作用,忽略了交互作用会导致我们得到错误的结论。

在设计单自变量实验过程中,你会将其他环境因素作为控制变量,并认为实验结果会受到变量水平影响,此时你就是在担心是否存在交互作用。"它或许依赖于"在不断的告诫你。比如,有领袖的群体解决问题是否更快? 这或许会依赖于群体规模。字体是否会影响阅读速度? 这或许会依赖于读者的年龄。看暴力电视节目是否会影响儿童的攻击性? 这或许依赖于他们看了多少。一旦你知道你所得到的实验结果还可能依赖于其他因素,如果将这些环境因素作为了控制变量或随机变量,那么,就存在犯错误的危险。看图 9-7 所示的实验结果,我们可以说,如果不做因素实验,单自变量的实验已经非常完美了;如果我们将群体规模作为控制变量,只选择了 3 人的群体,我们会得到"问题解决的时间与是否存在领袖无关"的结论;但如果我们选择了 20 人的群体,我们又会得出是否有领袖会影响决策时间的结论。

如果我们将可能依赖的环境因素作为随机变量,情况也变好不了多少。领袖实验会产生图 9-7 中的结果,如果我们随机地在 3 和 20 之间选择群体规模,这会使我们低估领袖的作用。换句话说,我们会发现,领袖发挥的作用较小,因为在随机选择的群体规模中,领袖的作用会变得比较平均。假设潜在的交互作用是另外一种形式,如图 9-8 所示,此时如果群体规模随机变量,我们仍会将领袖的作用在群体规模上进行平均,会错误地得出有无领袖对问题解决没有影响的结论。从这些讨论中,你也许开始认识到,为什么因素设计在心理学实验中使用得如此广泛。它是唯一能

够让我们考察变量间交互作用的设计方式(对于如何解释交互作用,请参看第12章)。

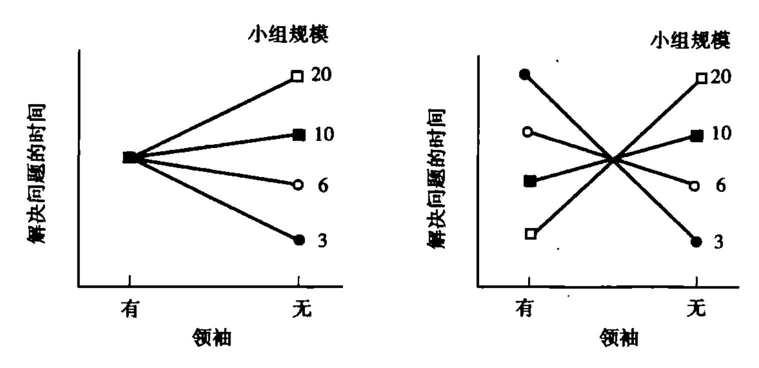


图9-8 领袖和小组规模之间可能存在的两种交互作用。在这两种情况,将小组规模变成随机变量都可能消除领袖的效应。

在第2章中,我们已经发现,当将某个环境因素作为随机变量时,实验结果的概括性会增加,但是准确性却降低。另一方面,如果将某个环境因素作为控制变量,则会增加实验结果的准确性,但又降低了概括性。因素实验给了我们第三种选择,我们可以将环境因素作为另一个自变量,从而同时增加了实验结果的准确性与概括性。我们可以把实验结果推广到更大范围的环境中,因为此时将许多环境因素都作为实验变量来处理,我们可以非常精确地知道这些因素的某种水平对因变量的作用结果。因此,尽管每次我们选择将某个环境因素作为实验变量,实验会变得非常复杂,但我们可以更清楚地了解所有可能的未知世界。

因素设计的第三个优点是统计上的优势,回忆第8章中提到的内容。最有推断力的统计检验是比较自变量水平之间的差异与数据本身的变异哪个更大。如果自变量水平间的差异非常大或者数据本身的变异比较小,那么,这种差异更有可能是显著的。在因素设计中,当某个能够增加数据变异的环境因素成为了实验变量时,数据的无关变异会降低。因此,成为实验变量的环境因素越多,数据的无关变异越小;无关变异越小,则差异显著的可能性就越大。

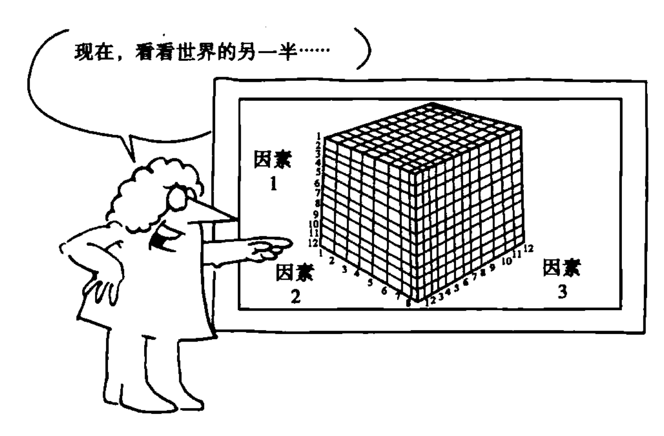
缺 点

因素设计尽管很好,但这种设计方式仍然存在一些缺点。因素设计主要的缺点是耗时、成本高。正如第2章所述,假如你为空军某个部门工作,同事是设计新型飞机驾驶舱的工程师。因为你是心理学家,对人非常了解,他们希望你能够告诉他们如何设计仪表和控制装置,并如何将它们安装在机舱内。

你知道一些因素可能与其他因素产生交互作用,因此,选择了一个因素设计实验。例如,你知道速度仪的位置会影响到高度仪的位置,你选择的第一个因素是高度仪指针的长度,你发现,当前使用4种标准长度的指针,因此,这一因素有4个水平。你还有5种放置高度仪的可选地方,因此,你把高度仪的位置作为第二个因素,它有5个水平。你的第三个因素是速度仪的大小,有3种水平。同样速度仪可选的

位置有 6 种，这样就产生了第四个因素。第五个因素是操纵杆的把手①，它有 4 种可能的直径和 5 种可能的长度。接下来开始考虑机舱设计的重要变量，我们已经形成了 4 ×5 ×3 ×6 ×4 ×5 的实验设计，它包括 7 200 个单元②。如果我们每个单元分配 10 人，那么，我们需要的人数就远远超过了空军飞行员的数量！

你会发现，当你将一个因素加入到因素设计中时，它会成倍地增加单元数量。按照这种比率，因素设计的规模会很快变得难以处理。因为每个增加的单元都需要更多的时间和精力，因此，因素或水平的数量一定要符合实际情况。



如果你没有可以完成如此庞大的因素实验的资源，那么，怎样才能得到一般的结论呢？解决方法之一是做几个小实验。例如，上面的实验中我们可以分为 4 ×5，3 ×6 和 4 ×5 的实验。这种解决办法的问题是，你假设出现在独立实验中的自变量（如高度仪的位置和速度仪的尺寸）不会相互影响，但在没有将这些自变量合并在一个实验中之前，你没有办法证实这种假设的正确性。然而，这正是绝大多数心理学家在现实世界中寻找答案的方法，在本章的后面将详细讨论这种做一系列小实验的策略。

还有第二种，也是解决庞大实验更成熟的方法，被称为反应—表面法（response-surface methodology）（Clark & Williges，1973；Meyers，1971）。该方法可以使你在不必将每个单元的数据点填满的情况下，确定出在因素设计中的哪个位置上因变量具有最大值或最小值。为了做到这一点，必须假设不会出现一些复杂的交互作用，这通常也是正确的假设。如何具体使用反应—表面法超出了初学者的需要，你现在仅仅需要知道，当你将来需要这种技术时，便可以用它即可。如果你需要使用这种设计，本章最后的参考是个好入门知识。

因素设计的第二个潜在困难是对实验结果的解释。绝大多数因素设计或两个以上水平的因素实验的统计分析方法是方差分析。它需要你对数据变异的类型作某种假设。其中一个假设是变异，是正态分布的。如果你的数据中潜在的变异不是接近正态分布，那么，进行方差分析是不合适的③。不幸的是，通常直到实验做完以

---

① 操纵杆是飞行器上的控制装置。

② 虽然前面已经说过，× 读作 by，所以，这是个 4 by 5 by 3 by 6 by 4 by 5 的实验，在确定单元的数量时，× 可以当成乘号使用。

③ Bradley（1968）在他的 *Distribution - Free Stastistical Test* 一书中详细地讨论了，当不满足这个假设时你可能会犯的错误。

后你才知道是否满足这个假设。如果是这样,那就比较糟糕了;因为其他统计方法无法分析复杂的交互作用。这种情况下,你只能不情愿地使用有争议的统计方法,或完全不进行统计分析。幸运的是,许多因素实验的分布是接近正态的,因此,你可以直接进行方差分析。(我们在附录 A 中详细讨论方差分析)

即使数据满足统计分析的假设,解释复杂的因素设计实验的结果通常也是非常困难的。到目前为止,我们提到的交互作用都是两维的(two-way),即一个自变量对因变量的效应依赖于第二个自变量的水平变化。然而,在第 12 章中还会讨论 3 维交互作用,即两维交互作用的类型和大小依赖于第三个因素。例如,领导力与群体规模交互作用,只是影响男性成员。有时,还会遇到 4 维、5 维的交互作用,那么,结果的解释变得更加困难了。

综上所述,与单一自变量实验相比,因素设计实验存在许多优势,它可以使你考察交互作用,在降低无关变异时体现了统计优势,使你在不降低实验准确性的同时,提高实验结果的可推广性。然而,你得到上述优势的代价是付出更多的时间和精力以及实验结果的解释困难。是否存在可以利用多变量设计实验的优点,同时又可避免这些难题呢?答案是肯定的(请继续往下看)。

聚合序列设计

许多杂志中都有报告一系列实验结果的文章,因为如今的许多实验都使用聚合序列(Converging Series)设计。聚合序列设计是指任何一种渐进地寻找解决问题的方法,而不是追求一下子就能解决问题。大多数序列设计实验是由多个单一自变量实验或比较小的因素设计实验组成。

序列设计的类型之一是,我们有一个待解决的应用问题,如果只做一个因素设计实验,则过于庞大,比如机舱设计实验。这种情况下,我们可以连续做一系列较小的因素设计实验,因为我们并不太关心非常复杂的交互作用(3 维、4 维或更多维度)。一旦我们找到了某个因素的最佳水平,在接下来的实验中,我们就可以将这个因素作为控制变量。因此,我们可以操控其他因素,直到成功地操控了所有预期能对行为反应产生影响的自变量。通过这样的方式,我们就可以逐渐接近整个问题的最佳解决办法。

聚合操作

与解决某一实际问题的实验设计相比,更令人振奋的是,存在一种聚合序列设计,它通过解释某观察行为的实验假设聚合的方式,去检验某个心理学理论,这种实验方式被称为聚合操作法(Garner, Hake & Eriksen, 1956)。我们在开始进行序列设计实验时,怀有多种能解释某种被观察行为的假设。每做一个实验就可以排除一个或更多的初步假设,直到最后只剩下一种能够解释数据为止。

多种可能的假设

实验一之前

词的特性
知觉防御
无意反应抑制
有意反应抑制

实验一之后

知觉防御
无意反应抑制
有意反应抑制

实验二之后

无意反应抑制
有意反应抑制

实验三之后

有意反应抑制

聚合操作

聚合操作

图 9-9 聚合操作实验步骤

为了详细说明聚合操作技术，请看下面的实验。我们想考察一下人们对粗俗词的识别时间是否长于非粗俗词。假设实验者使用速示仪呈现词汇，速示仪是一种常用于显示视觉刺激的设备，它可以很好地控制呈现刺激的时间。主试呈现4个单词，两个粗俗词，两个非粗俗词，要求被试一旦认出了词，就大声读出来。结果表明，被试需要更多的时间来报告粗俗词。因此，可以得出结论认为，此结果支持人们会无意识地抑制粗俗材料的假设。这种知觉防御假设认为，要克服这种抑制需要更多的时间。

作为优秀的实验者，你会考虑多种可以解释相同实验结果的其他假设。首先，词汇本身的特性会使非粗俗词更容易阅读；其次，被试可能对4个词的识别程度相同，但可能不自觉地抑制了对粗俗词的反应；第三，被试可能已经意识到了这些词汇，也知道对它们如何作出反应，但是有意在控制反应，直到他们十分确信正确时才作反应。因此，我们可以拥有至少4种能解释实验结果的假设，如图9-9所示。接下来我们需要做一系列实验才能聚合到其中的一个假设上面，而排除其余的假设。

你做的第一个实验可能是区分词的特性假设和其他3个假设。你可以使用4个不同的词重复最初的实验，如果再次发现粗俗词的识别时间长，那么，就可以将词的特性假设排除①。如果朗读粗俗词的时间并不长，那么，你就更有理由接受词的特性假设②。

假如已经排除了词的特性假设，你仍旧需要区分其余3种假设。在实验二中，我们试图确定，被试是否对粗俗词的知觉时间比他们报告的要短。我们还记得，皮肤电（GSR）可以显示出一个人对刺激的情绪反应。因此，我们决定在呈现词汇的同时，测量被试的皮肤电反应，从而发现词汇需要呈现多久才能被识别。皮肤电可以显示被试是否能够识别词汇，即使他们可能会有意识地或无意识地抑制自己的反应。

如果你发现，直到被试开始报告粗俗词时皮肤电才发生变化，知觉防御假设便得到了支持。如果对粗俗词和非粗俗词的皮肤电反应相同，则再去验证另外两个假设哪个正确即可。

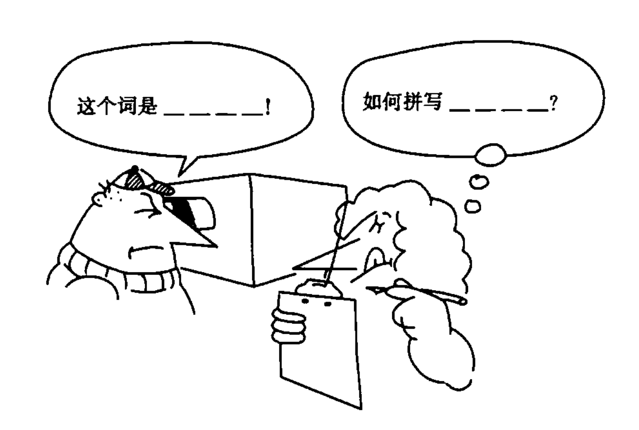
为了区分是否存在有意识的和无意识的反应抑制，你可以使用一种能够引起被试有意识地改变抑制程度的操作。你会预测，主试与被试的性别不同会比性别相同时产生更多的有意反应抑制。因此，实验三是，当主试与被试性别相同时，两类词汇识别的时间差别是否比较小。如果是，则支持有意反应抑制假设；如果不是，则可能存在无意反应抑制。

你可以看到聚合操作是如何让我们排除假设，直到最后剩下一个。向某个假设进行集中的操作是不断变化的，我们也可以选择其他操作，但如果操作背后的假设

---

① 实际上，仅做一个实验很少能够排除一个假设。例如，我们可能不那么幸运地选了其他两个粗俗词，其朗读时间仍旧比较长；或者我们无法仔细考虑其他的假设，如这种效应可能是由于粗俗词比非粗俗词的使用频率低，而我们对高频词的识别更快。如果十分肯定地排除这个假设，聚合操作必须不受任何其他可能操作的影响。词汇的改变，不能让词频完全独立于词粗俗与否，因此，我们不能排除词频假设。

② 这个句子是经过仔细斟酌的，因为我们无法真正提供支持字体假设的有力证据。在实验心理学中，我们设计实验的目的是，揭示由于对自变量的控制所产生的因变量的变化。如果自变量没有引起因变量的变化，并不能强有力地支持自变量不会产生变化的结论。因为有许多原因都会导致没有差别，如被试没有根据指导语的要求进行反应、睡着了或死掉了等。



是正确的，所有其他的操作都应聚合到这一假设上来。如果每次都有一个操作向某个假设聚合，则会增加我们支持那个假设的信心。

实际上，上面的讨论有些理想化。在完成一系列的聚合实验并详细检验每一个可能的假设和操作之前，你无法安心坐下做实验。像绝大多数的实验者那样，每次只能完成一个实验，只有在看到一个实验结果之后，才可以决定如何接着去做更靠近真理的实验操作。

如果你已经完成了序列设计中的多个实验，你会发现，假设的数量在逐渐增加，而不是减少。虽然你可以排除一些旧假设，随着对研究问题更加深入地理解，其他新的假设会不断涌现。在这点上，你看起来更像在完成分散的序列实验，而不是聚合的序列实验。然而，实际上，你仍在聚合，只是潜在的假设比最初想象的增加了许多而已。

优　点

通过上述讨论，我们可以明显地看出聚合序列实验法的大多数优点。在大型的因素实验中灵活性相对较小。在大型的因素设计实验的实验之前，必须确定选择相应的因素以及每个因素都应该有哪些水平，你的操作都会被限制在这些前提条件之中。一个错误的选择会耗费大量时间和经费。然而，在聚合序列实验中，你的选择会更多一些。在每个种情况下，你都可以选择一个新的自变量或水平。因为没有把大量的时间浪费在考察对因变量作用不大的因素和水平上，因此，效率更高。聚合序列设计还具有可重复的优点，即每次重复实验都能得到类似的实验结果，在科学领域中结果的可重复性是人们所推崇的。在你完成了粗俗词实验中的三个实验后，如果每次都重复得到一致的结果，三次实验都显示粗俗词的反应时间更长，这说明，实验结果是可信的。

缺　点

聚合序列实验设计也具有一些不足之处。如果将变量分别控制在不同的实验中，那么，要确定变量间如何相互作用是困难的，有时甚至是不可能的。在某些情况下，你可以将两个聚合序列设计中的实验合并在一起，而将其作为一个被试间的因素来设计实验。然而，如果你主要对交互作用感兴趣，那么，你就应该做因素设计实验。

第二个缺点是，当把序列设计中独立实验的结果进行比较时，你通常是在进行被试间比较，这将使它同时具有了被试间设计的所有缺点（见第8章）。

最后，当你在使用聚合序列设计时，在进行下一个实验之前，必须分析和解释上一个实验的结果，这种分析通常需要用几个星期，有时甚至数月才完成。因此，许多实验者都同时进行多个序列的实验，这样可以在做一个序列设计中的实验同时，还可以分析另一个序列设计中的实验结果。

综合聚合序列设计的优点和缺点，就可以很容易看出为什么这种设计方法最近如此受欢迎。聚合序列实验法提供了考察应用问题或基础研究问题的一种高效率和高灵活性的方法。

表9-1 总结了本章中已经讨论过的所有实验设计的优点和缺点，以供参考。

表9-1 两水平实验、多水平实验、因素实验聚合序列实验的优缺点

设计	优点	缺点
两水平实验	是确定变量是否存在效应的有效方法。 结果易于解释与分析。 适合于一些理论检验。 适合于应用研究中的比较。	无法推测函数的形状。 使用内推和外推都很危险。 难以检验复杂的理论。
多水平实验	可以对函数的形状进行推论。 自变量的范围并不重要。	需要大量的被试和时间。 平衡非常烦琐。 统计计算较为困难。
因素实验	可以分析交互作用。 因素的数目增加降低了变异性，因而增加了统计敏感性。 增加了结果的广泛性，但并没有降低精度。	随着因素数量增加，实验变得庞大。 统计分析变得更加困难。 有时，高阶的交互作用更难于解释。
聚合序列实验	比大型的因素实验更加灵活。 可以在内部进行重复。	难以获得交互作用。 实验之间的比较也是被试之间的比较，都与难度有关。 在进行下一个实验前，必须对前一个实验进行分析。

小　结

一旦你确定要研究一个值得探讨的问题，你就必须选择一种实验设计。最简单的设计方式是只有两个水平的单个自变量设计，它提供了快速确定自变量是否对因变量有作用的方法，并且实验结果容易解释和分析，它为一些基础问题或应用问题提供了所有的必要信息。然而，这种简单的实验不能回答实验关系的具体形式，因此，实验结果的内推和外推都存在风险。加入更多的自变量水平，可能使你能够更好地弄清自变量和因变量之间的函数关系，并且使得对自变量水平范围的选择变得不那么关键，这种多水平实验的缺点是，需要花费更多的时间和精力，并且实验结果

也更难以解释和分析。

最常用的多变量实验设计被称为因素设计,在这种设计方式中,将被称为因素的自变量合并在一起,因此,一种自变量的每个水平都会被组合在其他每个自变量的不同水平上。如果将被试内的因素与被试间的因素进行联合,这种设计就是混合因素设计。因素设计可以考察交互作用。每当增加一个因素时,虽然统计检验的灵活性降低了,但实验结果的可推广性和准确性都得到了提高。然而,大型的因素设计实验非常费时费力,如果这种设计变得非常庞大,那么,可以使用一系列小实验或使用反应—表面法。实验结果的解释也是个比较难的问题,当方差分析的统计假设无法满足时,尤为如此。

可以使用聚合序列设计来取代复杂的因素设计,这种设计可以让你使用聚合操作的方法。聚合操作法可以逐渐排除一些假设,直到最后只有一种可以解释实验结果的假设为止。聚合序列设计具有灵活性和可重复性的优势,然而,估计不同实验中因素之间的交互作用会变得比较困难。你必须将这些因素按照被试间设计的方式去操控,且必须将上一个实验结果彻底分析后才能进行下一个实验。

10

非实验研究设计

研究者在解释准实验结果时所面临的基本任务是，从每个处理组内单元之间最初无法比较的各种效应中找出处理效应。

T. D. COOK & D. T. CAMPBELL(1979)

对于研究者来说，不是在一千只老鼠身上花费一个小时，也不是在一百只老鼠身上花费十个小时，而是用一千个小时研究一只老鼠。

B. F. SKINNER(1966)

到本章为止,我们已经集中讨论了实验设计。但是,仅通过设计一个实验来回答某一特定研究问题通常还不够。本章主要讨论三种不太严格的实验设计方法。一是准实验设计(quasi-experimentation)。准实验设计遵循许多真实验设计的法则。但是,因为我们不可能将被试随机安排到自变量的不同水平上,所以,准实验设计必须最大程度上减少内在的无效来源;二是单被试和小样本基线设计。因为这种实验设计可以清晰预见每一个被试的结果,所以,我们能够建立一些规则,使得在没有控制组或被试内无法平衡的条件下也可以观察到效应;三是调查或问卷研究。它使用相关设计,而不是实验设计。

准实验(和非实验设计)

第 2 章曾经提到,分配条件的方法之一是将其转化为随机变量。第 1 章也强调了随机化的重要性。如果我们不能实现一个真正的随机化过程,那么,干扰条件可能会伴随自变量的不同水平造成系统的变化。换句话说,产生变量混淆的风险增加了。产生混淆的可能性无时无处不在,但是,我们必须对所有威胁内部效度的潜在因素保持警惕。这些因素已经在第 2 章讨论过,它们包括历史、成熟、选择、死亡、测验和统计回归。还记得这些因素吗?准实验设计的目的就是在无法对组别进行随机化处理的情况下,将这些潜在威胁降到最小。我们希望通过这种方法避免变量混淆。

举例说明这一问题。假设我们研究的问题是:每一节课结束时进行学习评价(短时、无学分的小测验)能否提高大学某一课程的专业测试成绩。参照基本的实验模式,我们至少使用自变量的两个水平——学习评价和无学习评价。有些事件将会成为控制变量,如教给两组被试同样的课程材料。但是,我们不能因为使用被试间设计而挑选学生;被试内设计也不可行,因为我们无法排除学生们第一次学习材料获得经验的影响。因此,我们必须随机安排每一组学生。理论上,我们需要将这所大学中没有参加该课程的所有学生姓名进行随机编码,从中挑选 100 位学生随机分配到有学习评价的班级和无学习评价的班级。

然而,现实情况是学校允许学生们自由选择课程、班级。研究者不得不采用两个已经存在的班级,例如,一个早上上课的班级和一个下午上课的班级。然后,将他们安排到自变量的不同水平。你可以设想,选择早上课程和下午课程的学生之间存在各种差异吗?你能想象这些学生有差异的维度吗?这些维度可能与课堂成绩相关。即使两个班级都是早上上课,一些人可能选择星期一、星期三和星期五,而另一些人可能选择星期二和星期四,星期二和星期四的课程时间可能更长些。你能想象不同的上课时间和课程持续时间对学生成绩的影响吗?另外,如果老师每一学期或每一学年只教授特定课程的某一章,那么,春季班和秋季班的学生之间、本学年和下学年的学生之间都会存在维度上的差异,你能想象出这些情况吗?因此,当使用控制和随机化方法避免有些事件成为混淆变量时,我们无法控制被试的分配。

在大多数应用领域中,我们无法对被试进行随机安排,因此,有时只能使用准实验设计。准实验设计可以将各种影响内部效度的因素降到最低程度,在有些情况下至少可以确定这些因素。下面我们讨论的每一种设计都可能遭遇这些威胁,但没有

人能够给我们一个确定的办法消除它们。就准实验设计而言，采用 Cook 和 Campbell(1979)著作中的符号系统描述这些不同设计的特征非常方便。其中，X 代表了自变量的特定水平(也称处理)，O 代表了一个观测，即因变量的测量。下标 1 到 n 代表了处理的顺序($X_1 \ldots X_n$)或者观测的次数($O_1 \ldots O_n$)。实验组之间的破折线表明，被试并非随机选择。

非实验设计

单组后测设计

如果你只是测量一组被试在某个自变量的一个水平上的行为，那么，这就是一个单组后测设计。用符号系统表示如下：

X　　O

如果没有其他信息补充这一结果，这种设计对于判断处理效果一定是无用的。

例如，一家电视网络公司正在播出一档关于大屠杀的节目(X)。你想知道这档节目如何影响人们对该事件的关注？通过对一组被试的问卷调查，你发现 76% 的人正关注大屠杀期间发生的事。从这一结果来看，你认为电视节目产生了什么作用？它引起人们对这一事件的关注程度增强，还是降低？如果你不知道电视节目播放之前人们的意识水平，或者不了解另一个没有观看节目的对等组的水平，那么，调查结果对于回答这些问题毫无用处。

单组后测设计与第 1 章讨论的个案研究方法十分相似。但是，相比较而言，个案研究显得更加有用。在个案研究中，研究者知道大量被观察行为的背景信息。所以，即使没有直接测量观察前的行为，也可以将它们推断出来。而且，个案研究观察的行为通常不止一种，这些行为将形成一个模式，它所提供的信息远比实验室情境中测量单个因变量所得到的信息多得多。

不对等组后测设计

如果在单组后测设计的基础上，增加一个不对等组的后测，我们可以得到以下设计：

X　　O
- - - - - - - - - - -
　　　O

不对等组的选择机制与处理组的选择机制完全不同。

假设在大屠杀节目的例子中，我们发现，因为迈阿密当地足球队正在比赛，所以没有播放这档电视节目。此时，我们从迈阿密随机选择样本作为不对等组，对他们进行问卷调查。结果显示，两组之间存在差异。那么，我们能将这一差异归咎于电视节目吗？考虑到迈阿密有大量犹太人，你是否考虑过犹太人可能影响你对大屠杀的关注呢？

不对等组后测实验设计的基本问题是，任何观察到的差异有可能源于组别间的选择差异，也有可能是源于实验处理的差异。总之，组别越对等，结果越可信。

在没有正式前测的情况下，增加结果可信度的方法是进行非正式前测，使得两

组被试可比较。前测信息越有效，它和因变量的关联度越高。因此，我们可以根据年龄、性别、社会阶层、民族和信仰等信息比较两个样本。通过这种比较，我们可以确定两组是否对等。但是，即使这样，不对等组后测实验设计仍然缺乏说服力，我们必须慎重解释结果。

单组前—后测设计

如果我们在单组后测设计的基础上，增加一个前测，就形成了单组前—后测设计：

$$\overline{\underline{O_1 \quad X \quad O_2}}$$

这种设计被广泛应用，是对不对等组设计的一种改进。与被试内设计相似，对同一组被试实施两种观测，可以使选择因素的干扰降到最小。但是，就被试内设计而言，刺激呈现顺序的平衡、两种观测的衔接紧密程序等都会降低威胁内部效度的其他因素的干扰，而单组前—后测设计却做不到。

仍然以大屠杀节目为例。你认为给予一个前测（如询问人们对于该事件的关注程度）会对后测（如评价人们关于大屠杀的关注）产生什么效应？测验影响内部效度的可能性依然存在。即使在实验处理前将测验的威胁降到最低，如两次测验间隔一年，你也会陷入其他的威胁当中。历史因素同样产生威胁，因为除电视节目外，其他一些与大屠杀相关的事件（如抓获战犯）也可能改变被试的关注程度。如果被试是学生，成熟也会影响内部效度。当我们使用前测去选择实验组时，回归也会产生问题。因此，我们在使用前测设计解决被试的选择问题时，依然要慎重解释结果，因为其他威胁内部效度的因素同样存在。

准实验设计

上面论述的3种设计被称为非实验设计，因为它们无法估计许多威胁内部效度的因素。本节讨论的是准实验设计，虽然准实验设计达不到真实验设计的严格要求，却能够估计许多威胁内部效度的因素。本书并不能穷尽所有准实验设计，只是提到两种主要的设计类型。具体内容请参照 Cook 和 Campbell（1979）或 Shadish，Cook 和 Campbell（2002）的著作。

不对等控制组前—后测设计

第一种准实验设计使用一个不施加处理的不对等控制组和一个实验组。两组都给予前测和后测，用符号系统表示如下：

$$\begin{array}{ccc} \hline O_1 & X & O_2 \\ \hdashline O_1 & & O_2 \\ \hline \end{array}$$

该实验设计被广泛应用在社会科学领域研究中，因为它可以对大多数干扰内部效度的因素进行评估。

在一定程度上，我们所担心的那些威胁因素依赖于特定的实验结果。如果两组前测分数无差别，我们可以在一定程度上认为，两组是对等的，选择和回归因素产生

干扰的可能性最小。如果控制组在前、后测中的分数相同,则历史和成熟因素的威胁性最小。因为两组接受了同样的测验,所以,测验的差异效应也最小。但是,如果两组被试在前测、后测中亡失的数目不同,死亡将成为一个问题。重要的是,该实验设计可以评估这一影响因素。对于该设计而言,最可能存在的严重问题是与选择的交互作用。虽然两组的前测分数相等表明与选择的交互作用已经降低,但依然可能存在。例如,学校 A 正接受特定的实验处理,学校 B 没有接受处理。此时,如果学校 A 聘请了一位新校长,实施新的教师评价标准。那么,历史和选择的交互作用则会威胁我们的结论。

当两组的前测分数差异很大时,我们应该更加关注与选择的交互作用。举例说明:假设我们想知道,对装配线工人实施计件工资是否会增加生产效率?我们征集志愿者,降低他们的工资,却按计件付给他们额外的钱。前测结果显示,志愿者的生产效率更高。但是,当我们比较前测和后测差异大小时,发现后测差异更大:两组都提高了产量,但计件组提高更多。我们依此断定,计件工资能够提高生产效率。我们的推断正确吗?

生产效率存在前测差异:实验组的前测效率更高,成长速度更快(通过学习变得更有经验)。工人的生产技能缺乏稳定性,但会越用越熟练,因为控制组的技能也得到了提高。每一个人都在提高,所以,工人的技能越好,提高得就越快。不对等控制组前—后测设计并不能估计成熟—选择交互作用的大小。我们或许可以通过对实验组前测的细分解释这一效应。也就是说,我们期望得到两个实验组,即能力弱的、提高更慢的工人和能力强的、提高更快的工人。但是,这样一来,我们将会得到一个完全不同的实验设计。关键问题在于,即使采用不对等控制组前—后测设计,也仍然会面临一些威胁,如与选择的交互作用。

不对等控制组前—后测设计的变式。我并不想细究不对等控制组设计的每一个变式,只想提及一些可能的情况。在某些情况下,当前测和后测使用相同的测验不太可能或不切实际时,我们会使用一种近似的前测。也就是说,前测采用的一些变量与后测相关。例如,你想评价一种新的代数教学法的效应。首先,选择一个班级施以新的教学方法,而另一个班级保持原有的传统教学方法。接着,我们不是采用前测来评定每个班级在实行新教学法之前的成绩,而是采用一个接近前测评价的一般数学能力测验。

在不可能实施前测的情况下使用近似前测。例如,当处理包含的一些无法预测的历史事件影响了一部分人,此时采用近似的前测。或者,即使有可能进行前测,但测验会威胁到内部效度,我们也可以采用近似的前测代替前测。此外,若研究期待有新的效应,前测和后测使用相同的测验也是毫无意义的。举个例子,在两个班级未学心理学课程之前,对他们进行期末考试,这是毫无意义的。

如果测验威胁内部效度,我们将使用不同的前测和后测样本。在这种情况下,不是对每组的单个样本同时进行前测和后测,而是每组分别抽取两个样本:一个样本接受前测,另一个样本接受后测。例如,在一个班级实行新的教育计划,另一班级不实行。将这两个班级随机细分,每个班级都有一半人接受前测,另一半人接受后测。该设计有一个明显的弱点,即它完全取决于前测组和后测组的可比性。如果两组的差异维度与实验处理相关,设计的弱点就更大了。

增强不对等控制组前—后测设计适用性的另一种方法是,按多个时间间隔增加前测。增加前测有助于我们估计两种影响因素的效应。还记得我们前面讨论的“有能力的人怎样得到更多的能力”吗？它如何产生成熟—选择交互作用？一方面,如果我们很早就实施了前测,那么,我们就能够判断测验分数是否落到每一组的趋势线上。如果落到了趋势线上,我们就可断定,是成熟—选择的交互作用而不是处理导致了后测差异。也就是说,两个前测已经出现了成熟趋势,后测仅仅只是趋势的延续。另一方面,增加前测有助于我们估计统计回归效应。如果根据第一次前测来选择不同组,回归效应就会反映在第二次前测和后测的分数上。

还有一个很少使用的变式,程序如下:一个前测,施加处理,一个后测,去掉处理,再一个后测。如果接着恢复处理、给予另一个测验,如此延续下去,那么,该实验设计便可以无限扩展。这种变式与我们将要讨论的基线设计很相似。但是,基线实验设计的被试很少,单独考察每个数据,通常不进行统计分析。

通常情况下,第一组接受处理的目的是期望因变量在某个方向上发生变化;而对第二组施加处理则是期望因变量发生相反的效应。例如,在两组工人中,一组按计时工资,一组按计件工资。对他们施加的处理是,使得一组全部是计时工资,第二组全部是计件工资。如果我们的实验假设是:计件工资增加生产效率。结果应该是:第一组的生产效率降低,第二组的生产效率增加。结果支持我们的预测,也就强烈支持了我们的假设。

至此,我们已经讨论了不对等控制组前—后测设计的一些变式,具体程序参见表 10-1。当然,其他的变式也可能存在,本章最后也列举一些。

表 10-1 实施各种不对等控制组前—后测设计的程序

	时间 1	时间 2	时间 3	时间 4
基本的不对等控制组前—后测设计	测量组 1 测量组 2	施加处理 不施加处理	测量组 1 测量组 2	
接近前测	接近测量组 1 接近测量组 2	施加处理 不施加处理	测量组 1 测量组 2	
分离的前测和后测样本	测量组 1 的前一半 测量组 2 的前一半	施加处理 不施加处理	测量组 1 的后一半 测量组 2 的后一半	
多时间间隔的前测观察	测量组 1 测量组 2	测量组 1 测量组 2	施加处理 不施加处理	测量组 1 测量组 2

注:组 1 和组 2 的被试不是随机分配到处理和非处理条件下的,因此,是不对等的。

间断时间序列设计

准实验设计的第二种主要类型是间断时间序列设计。其中,单组在施加处理前被测量多次,施加处理后也被多次测量。用符号系统表示如下:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad X \quad O_6 \quad O_7 \quad O_8 \quad O_9 \quad O_{10}$$

如何解释结果呢？最简单的方法是观察实验处理后发生的变化是否持久不变。

例如，某工厂采取一种新的工资方案后发现，工人的生产效率立刻提高了10%。如果这一变化在整个研究过程中得以保持，研究者就可以确信：新的工资方案提高了生产效率。但是，即使我们得到了一个理想的结果，也仍需要谨慎各种可能的威胁，如历史和死亡。一些历史事件可能与实验处理相互影响；再者，施加处理的那一刻，一些不明事件可能引起被试的大量亡失。



间断时间序列

间断时间序列设计还可以排除或估计其他一些影响内部效度的因素。例如，因为整个实验过程采用的组别相同，所以，与选择的交互作用可以排除；测量和统计回归的效应也在施加处理前消失了。通常，因为成熟效应的作用不大，我们也可以排除它。因此，我们期望的结果是一条趋势线，而不是一种非连续的变化。

然而，当因变量的变化是延迟的、暂时的，或者反映在增长或降低趋势斜线上，而不是整个平滑水平线上，我们就不能确定地作出结论。在这种情况下，更高级的统计技术有时可以帮助我们分析处理效应。

间断时间序列设计的变式。第一种变式是增加不对等、无处理控制组的时间序列设计。其中，这个不对等组在每个观测间隔也被测量，整个序列不施加实验处理。增加控制组可以对历史效应的影响进行估计，因为两组被试受到相同历史事件的影响。如果两组被试的选择方式不同，历史—选择之间的交互作用将会产生影响。但是，这种影响只有在一个特定的历史事件与处理同步发生的情况下才会出现，而且只限于处理组。

第二种变式是去除处理的间断时间序列设计。此时，处理效应是可逆的。当基本的时间序列设计完成之后，去掉处理，进行另一个序列观察。这一设计实际上结合了两种基本的时间序列设计，一个序列实施处理，另一个序列去掉处理，它们之间有所重叠。事实上，你可以根据需要多次增加和取消处理，产生多次重复①。每一

① 根据伦理道德原则，如果已经发现处理是有益的，就必须结束处理序列。

次重复都会增加你对处理效应的判断的信度。此外,这一设计与后面讨论的基线设计相似。

重复的另一种方法是使用不对等组,即同时在两组观察序列的不同时间点施加处理。这一设计被称为转换重复的时间序列设计。该设计可以平衡或估计大多数影响内部效度的因素,如历史和成熟。此外,通过在来自不同总体的样本中嵌入重复,该设计提高了实验结论的内部效度。表10-2总结了这些间断时间序列设计。

表10-2 实施各种间断时间序列设计的程序

时间1	时间2	时间3	时间4	时间5	时间6	时间7
			基本的间断时间序列设计			
测量组1	测量组1	测量组1	施加处理	测量组1	测量组1	测量组1
			增加不对等、无处理控制组的时间序列设计			
测量组1	测量组1	测量组1	施加处理	测量组1	测量组1	测量组1
测量组2	测量组2	测量组2	不施加处理	测量组2	测量组2	测量组2
			去除处理			
测量组1	测量组1	施加处理	测量组1	测量组1	去掉处理	测量组1
			转换重复			
测量组1	施加处理	测量组1	测量组1	测量组1	测量组1	测量组1
测量组2	测量组2	测量组2	施加处理	测量组2	测量组2	测量组2
测量组3	测量组3	测量组3	测量组3	测量组3	施加处理	测量组3

准实验设计的统计分析

过去几年中,准实验数据的统计分析技术提高了很多。一些相当高级的统计方法超出了本书的范围,但我们可以参考本章最后列举的一些书目。虽然,一些准实验设计与即将讨论的基线设计十分相似。但是,基线设计的被试很少,不可能进行统计分析;而准实验数据采用的统计分析技术与真实验设计同样严格。准实验已经不再受统计分析的制约。

优 点

准实验设计的最大优点是允许我们做一些以前不可能做的研究。当心理学家们对社会问题、临床评估和教育计划感兴趣,并希望在实际生活中考察这些问题时,准实验设计提供了大量的新工具。研究者必须谨慎判断是否存在影响内部效度的因素,至少应该知道这些影响因素是什么。而准实验设计恰恰可以估计大多数的影响因素,并确定它们的性质。

缺 点

在我们大力宣扬准实验设计之前,也必须指出它同样存在一些局限性。即使我们倾尽全力,也会发现影响内部效度的因素依然存在。虽然,我们经常能够在它们发生的时候抓住它们,但是,也许某个因素就会导致结果无效。例如,在一个基本的

不对等控制组前—后测设计中，我们发现控制组的前后测成绩有差异，那么，解释实验组的任何变化就会存在困难。第二个问题是，与传统的实验设计相比，准实验设计显然更为复杂，研究者花费的时间和精力也更多。并非一种条件仅测量一次，所以需要花费大量的测量时间。另外，尽管准实验设计的统计检验现在是可能的，但是还存在困难，这也不属于基本课程的常规教学范围。

由于准实验设计的这些缺点，行为科学也经常倍受批评，或是因为对简单且不重要的问题作了彻底研究，或是因为对复杂且重要的问题作了不彻底的研究。但是，准实验设计的优势是可以对复杂且重要的问题作彻底研究。在解释准实验研究结果时，我们必须小心谨慎。与真实验设计和非实验设计相比，准实验设计的应用范围更广，考察更为广泛的问题的可能性更大。

单被试和小样本基线设计

单个数据 VS 群体数据

有些研究者认为，大多数心理学家的实验方法几乎都是一种误导，并且毫无根据。其中，反对呼声最大的研究者是 Sidman(1960)。他认为，传统的实验研究告诉我们关于被试行为的信息很少。实验通常告诉我们的是一些虚假的被试行为；这些行为并不能准确反映个体的任何真实情况。Sidman 指出，大多数实验使用不同组，并假设组中个体的行为与组的平均行为相似。但是实际上，在很多情况下，组中个体的行为与组的平均行为存在很大差异。

下面举例说明这一点。设计一个实验考察被试通过各种例子学习简单类推的速度！呈现第一项：edit 之于 tide 就好比 recap 之于______。答案是 pacer，因为类推规则是从后向前拼写。Tool 之于 loot 也遵循了这一规则。每位被试有 3 秒钟的时间去回答一个空。然后，呈现下一个项目。我们期望学习是以有或无的方式进行。也就是说，被试看到例子，突然在某些时刻喊出“Aha”或者“Eureka”，从那时起便领会了规则，每次答案都是正确的。

图 10-1 描绘了 10 个被试的结果；图 10-2 展示了该组的平均数曲线。通过比较可以发现，小组的曲线并不能代表图 10-1 中单个被试的曲线。根据小组曲线，我们认为，人们渐渐学会了解决问题。但是，实际上，单次试验中，每位被试都经历了“从解决不了任何项目到解决所有项目”的过程。

上述差异的存在使 Sidman 相信，小组的成绩并不能代表单个被试的执行过程。在第 8 章，Poulton(1973)对此问题提出了截然不同的观点。他认为，所有被试内设计基本上都存在缺陷，只有被试间设计使用不同组才有利于结果解释。心理学家们在实验设计时首先使用不同组，这是因为个体的行为如此多变，单个被试的变异性可能被其他人反方向的变化抵消。但是，Sidman 认为可变性并不是被试固有的特征，而是由于主试没有控制好所有影响被试的变量造成的。一旦主试合理控制了行为，他们就没有必要采用不同的组。而基线实验恰恰能证明主试获得了这种控制。

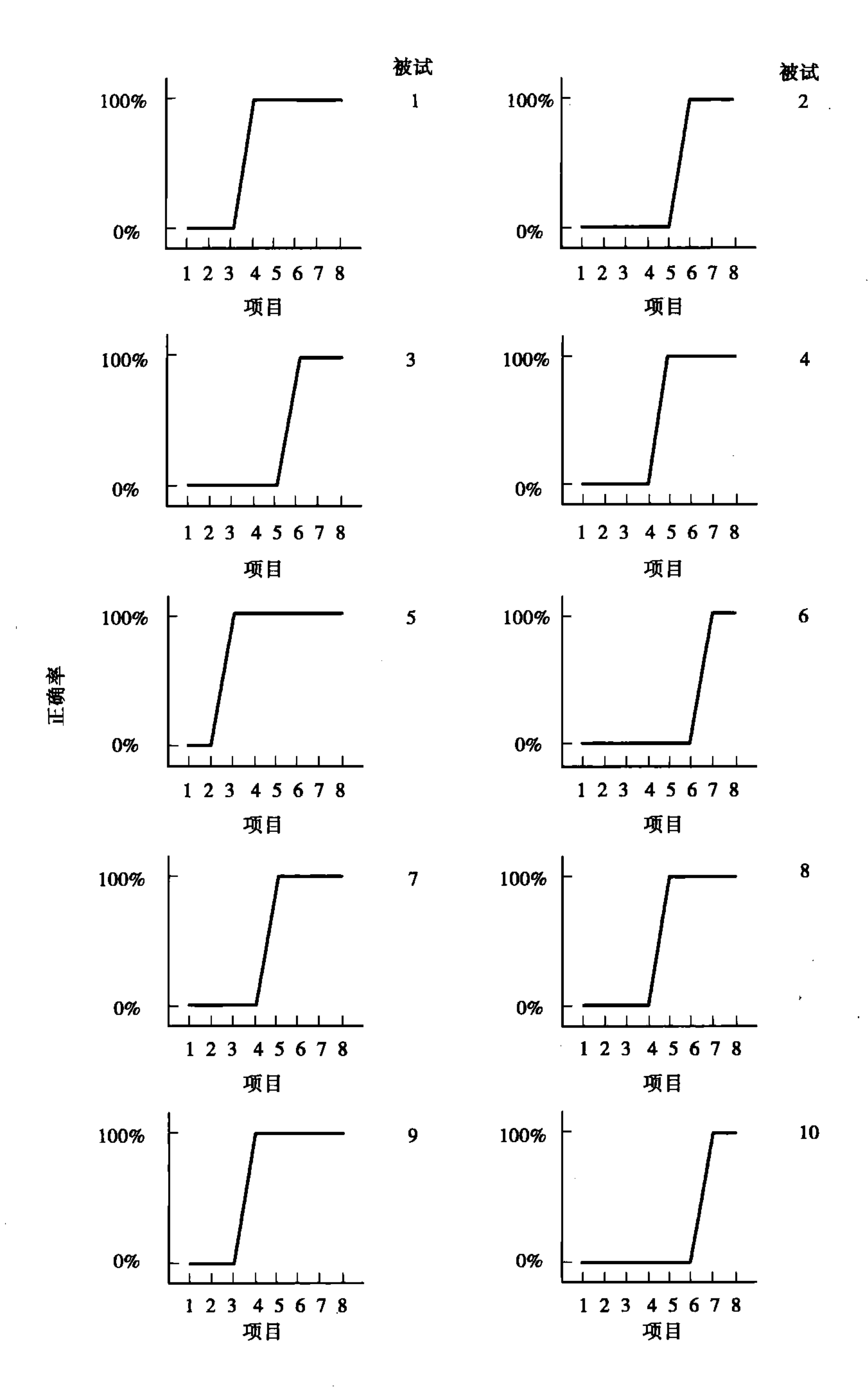


图 10-1　10 位被试在类推测量中的结果。一旦被试掌握了类推规则，剩下的所有项目答案都将是正确的。

基线程序

基线实验经常被提及是行为的实验分析。为解释 Sidman 的这种观点，我们设计了一个实验。惩罚能否改变脑损伤病人的行为？假设一位医疗专家正在治疗一

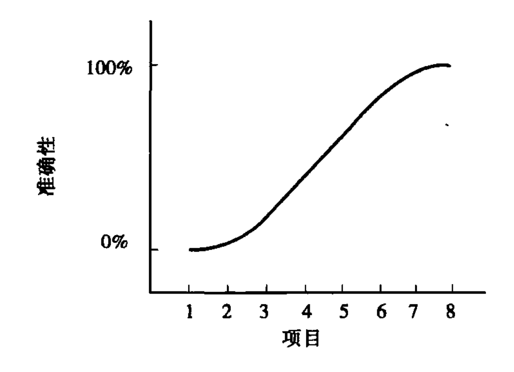


图 10-2 图 10-1 中所有被试的组曲线，组曲线并不能代表单个被试的行为。

位脑瘫病人，这位病人希望通过治疗提高自己的谈话技巧①。因为脑瘫病人通常不能控制自己的头动和目光接触，而目光交流是成功谈话的一个因素。因此，为增加病人目光接触的次数，医疗专家决定设计一个程序：每一次目光交流失败时，就给病人一次轻微电击作为惩罚。当然，这需要征求病人的同意。这位病人希望提高自己的社交技巧，同意了这一电击程序②。

基线实验的第一步是建立一个基线，也就是一个稳定的状态；在这一状态中，被试反应速度的变化很小。基线实验备受争议的一个问题是"反应速度变化究竟在什么程度上才算很小"。判断基线是否已经达到稳定状态的方法有很多。例如，设定一个统计标准：系列之间的反应速度变化不到 3%；或者对数据作一个简单的视觉检验，看是否有明显的波动或趋势。一旦基线被确定，主试就可以开始实验操作。

在我们的案例中，医生与病人每天有半小时的模拟谈话。然后，要求病人进行报告。谈话期间，若病人的眼睛没有保持接触，医生便会启动一个隐藏的开关。开关与时钟相连，以便计算每半个小时的谈话中病人眼睛接触的总时间。多次谈话之后，便可以建立一个稳定的基线水平，也就是说，每次谈话中的目光接触时间保持稳定。于是，医生开始启用电击程序。只要病人没有目光接触，医生就按下开关，病人的前臂会受到一个短暂的电击。然后，主试判断目光接触的次数是否从基线水平开始变化。



第一步是建立一个稳定状态……

① 我十分感谢 David A.（Sachs of Las Cruces, New Mexico）提供的被试案例。他设计了描述技术，尽管我报告的结果是虚构的。

② 因为道德伦理要求，必须征求病人的同意，尽管并没有充分的理由。

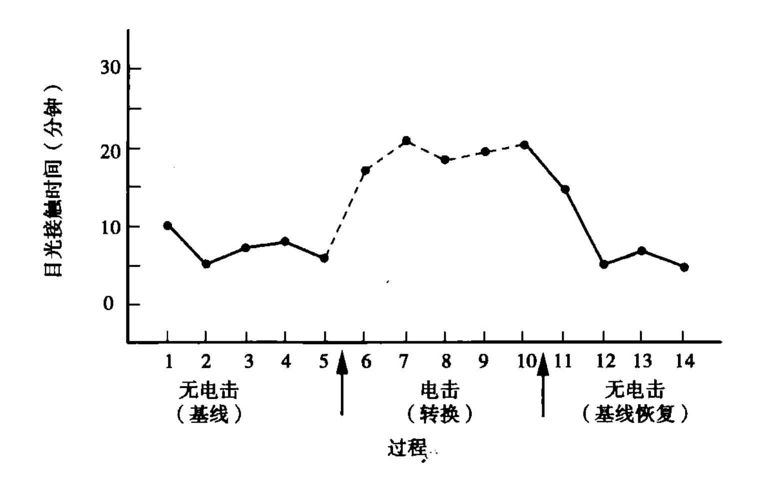


图 10-3　可能的实验结果。在 30 分钟的模拟谈话中，测量脑瘫患者的目光接触时间。前 5 个序列提供了一个基线，序列 6—10 执行了电击，序列 11—14 恢复到基线水平。

图 10-3 说明了实验可能的结果：5 个序列的谈话后，稳定的基线建立。第 6 个序列开始电击。一旦电击开始，病人的眼睛接触迅速增加。到第 10 个序列，眼睛接触已经达到了一个过渡稳定状态①，此时主试结束电击。到第 12—14 个序列，病人回到原始的基线水平。

主试必须执行以上描述的每一个操作，以实现一个真实的基线实验，即建立一个稳定的基线；然后采用实验操作建立一个转换稳定状态；最后，去掉实验操作，恢复到原来的基线水平，显示处理效应的可逆性。

这种方法的逻辑是，一旦建立了基线，在每一次实验操作中，就不会突然出现影响每次试验行为的无法控制的混淆变量。即使出现了无法控制的混淆变量，在同样的试验中，这些混淆变量对不连续的实验操作产生作用的可能性也非常小。

被试内重复是让相同被试在不同时间内多次重复实验程序，这可以增加实验结果的可信度。在我们的案例中，主试可以对被试进行第 15 个序列的电击，通过持续电击，直到获得稳定转换状态；结束电击后，又恢复到原来的基线。效应每被重复一次，我们就更加确信，行为的变化是由实验操作引起的，而不是无法控制的混淆变量引起。对于单被试设计而言，被试内重复—用其他一些被试重复实验—也能够增加结果的可信度。但是，我们仍然通过分析单个被试的数据，而不是小组的数据来评定结果。

优　点

基线实验的主要优点是，它提供了一种分析个体行为的有效工具。例如，如果图 10-3 的结果来自于实际数据，我会相信暂时的电击控制了目光交流。你也会相信的，不是吗？

实验结果也很容易解释。实际上，不使用统计检验也很容易解释基线实验。使

① 在临床上，转换稳定状态有时被称为处理状态或调整状态。

用基线设计的研究者们认为,如果你需要使用统计检验来让其他研究者相信,你发现的效应是真实存在的,而不是偶然的变化。这就意味着操作效应不够显著,你应该精炼技术,更好地控制行为(即消除意外的变异)。

在传统的分组实验中,如果每组都使用大量被试,你会发现效应在统计上是显著的,但是却没有任何实际意义。换句话说,你选择了一个影响行为的自变量,但与其他更重要的变量相比,它的效应很小。对于基线实验而言,这些问题是不存在的。因为其他更重要的自变量导致的变异性涵盖了这些真实却微小的效应。因此,基线实验可以保证任何效应都是有意义的。

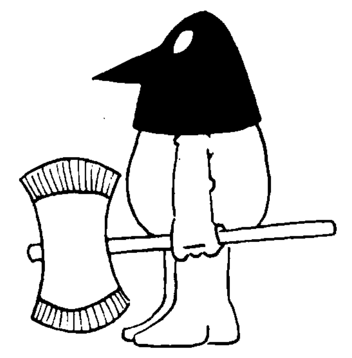
基线实验的第二个优点是灵活性。你可以灵活决定何时实施自变量一个水平的处理以及实施哪个水平。在进行一个标准实验前,研究者必须选择适当的试验呈现次数和自变量水平,这是大多数统计检验的需要;然后,收集每个自变量水平上大量相等的数据点。而采用基线设计的研究者可以在实验的任何一点收集当前这个水平上的更多数据;或者将其变成一个新的水平。例如,在我们的实验中,如果医生觉得有必要收集电击条件下更多的数据,那么,他可以在恢复到基线之前继续施加更多的电击序列。

医生也可以在实验开始后增加一个额外的自变量水平。假设行为的改变并不确定是由已选择的电击强度引起的。那么,在成绩恢复到基线并稳定之后,研究者可以在下一个试验组施加更强的电击。因此,基线设计不要求研究者预先决定自变量的水平。

除了容易解释结果、不使用统计检验、确保能够发现较大的效应和灵活性之外,基线实验还有一个优点是可以适用于单个被试。当医生只治疗一个脑瘫病人,或者实验者面对一个有某种不常见症状并需要训练的异常被试,或者不能使用传统的实验设计去研究某些能力时,基线设计是最佳的选择。

缺 点

尽管基线实验有许多优点,但是大多数研究者仍然坚持传统实验组设计,因为不需要满足基线实验的种种假设。例如,"实验效应可以逆转"这一假设要求被试在实验结束时回到原来的行为水平。前面的章节中提到,被试内设计需要平衡许多可能的顺序效应。基线实验是一种特殊的被试内设计,但是却无法对效应进行平衡。因此,任何一种系统地改变混淆变量的方法都会阻止我们恢复到原来的基线水平。除非在解除实验操作时行为恢复到以前的状态,否则,我们不会知道转换状态的行为是由于操作引起的还是一些混淆变量造成的。


有些效应是不可逆的。

因为这一原因,许多传统的心理学领域不能使用基线实验来研究,比如寿命、记忆和一些学习领域。这些领域在实验期间发生的大多数变化都是不可逆转的。(如,"现在忘记所有你学过的单词。")

基线设计的第二个缺点是,不能发现微小但又重要的效应。假设你在一家电话公司工作,你的任务是研究:与标准电话簿相比,电脑

查询系统是否可以节约操作员查询某个号码的时间。你决定采用基线设计，当每一通电话接进来时，你记录通话长度。你要求操作员首先使用标准电话簿，直到建立一个基线；然后，要求操作员使用电脑查询系统，由于信息被输进电脑，所以电脑能够提供号码；最后，你要求操作员重新使用电话簿。

图 10-4 呈现了一些可能的结果。直观上看不出来，使用电脑查询系统和标准电话簿之间的任何差异。换句话说，转换状态与基线看不出差异。但是，使用电脑查询系统的平均通话时间是 3 秒，比使用标准电话簿的时间短。如果我们已经完成了标准实验，统计检验会告诉我们差异显著。然而，这个效应重要吗？是的，它十分重要。因为使用电脑查询系统节省的每一秒通话时间累积起来，就可以为电话公司节省百万美元。这对于他们来说相当重要。

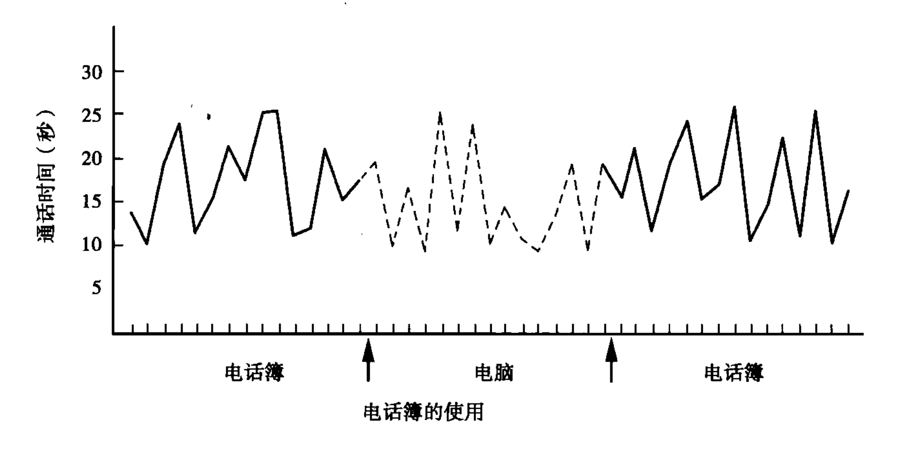


图 10-4　虚构的基线实验，测量操作员在使用标准电话簿和电脑系统两种条件下，每一次回复求助电话的时间。

支持基线设计的研究者认为，变异性是实验者的过错。例如，他们会在无菌实验室环境下研究老鼠。我经历了一个苦思冥想的困难时期，思考电话公司的科学家们怎样对测量的行为实行更好的控制。在本节的例子中，行为似乎在很大程度上被顾客而不是操作员驱动。在某些情况下，变异性还可能是由于情景设置的内在特征引起的。此时，微小却重要的效应可能被变异性掩盖，基线设计就不适合了。

基线实验的最后一个缺点是，很难判断研究所发现的任何效应的普遍性。因为个体对实验操作的反应各有不同，被试或许是一个十分古怪的人。采用额外的被试，或者在不同实验条件下使用相同或不同的被试，重复实验结果可以克服这一难题。但是，基线实验的传统是使用很少的被试。

在某些实验心理学领域，基线实验设计是一个很有价值的工具。历史上，有研究者在实验室使用这种设计，研究诸如简单学习等问题。一些心理学家认为，采用这种设计还可以研究其他重要的问题；如可控的卫生保健（Blampied，2000；Morgan & Morgan，2001）。如果研究者可以满足基线设计的这些假设，那么，该设计就为证实重要的实验操作效应提供了一个有效的途径。然而，这些假设如此苛刻，以至于基线设计很难广泛应用于众多实验心理学领域。

调查研究

调查研究的方法

调查是从样本中收集信息;然后,将这一信息概括到更广泛的人群之中。行为信息可以通过调查直接测量;不能被直接测量的信息,如意见、动机,甚至包括对未来的期望等,也可以采用调查法获得。在这些情况下,除了调查外,似乎没有其他更好的方法收集这些信息。因为调查不仅可以让你询问"人们做了什么",还可以询问"他们为什么做"。所以,调查法被社会科学广泛使用。如果学生们设计一项研究计划,首先映入脑海的通常是调查法。

收集调查信息的方法很多,最广泛使用的是邮件调查,直接将问卷邮寄给被试。(显然,调查研究者要求回答问题的人必须是接收邮件的人)邮件调查的优点有:抽样成本小,样本人群分布广泛。与面对面的交流相比,邮件调查难度较小,成本也较低。邮件调查可以保持匿名性和保护隐私,回答者更可能坦诚回答问题。当然,如果你有感兴趣群体的姓名和地址,你也可以发送邮件给这一群体中具有代表性的样本。

但是,邮件调查存在一个严重的问题是,人们有权利将调查邮件和其他邮寄宣传品一起当成垃圾扔掉。所以,邮件调查的回复率较低,有时甚至会低到20% ~ 30%。那么,为什么不多发送一些调查邮件,然后忽略那些未回复者呢?问题在于,可能会产生未回复偏差。我们已经在第2章讨论了未回复偏差,但在这里,它被称为选择——一种威胁内部效度的因素。就一项调查而言,你会谨慎选择样本,因为样本代表了大部分人群。你希望结果可以概括到大部分人群之中。但是,如果被试拒绝回答,样本成分就会发生改变。如果被试以随机的方式拒绝回答,样本仍然具有代表性,但这是不可能的。同时,研究者很难判断被试拒绝回答的原因,这些原因可能使样本发生系统偏差。或许有些人家境富有或教育程度较高,他们通常很忙,没有时间回复调查;或许有些人没有受过教育,无法阅读问卷;或许有些人在政治上是保守的,他们认为,你的调查侵犯了他们的隐私。真正的问题是,研究者并不知道,调查的未回复者与回复者之间的差异程度;也不知道样本系统偏差的程度。因此,研究者就不能将研究发现应用于大部分人群。

将未回复偏差降到最小的方法当然是获得高的回复率。Dillman(1978)提出了许多提高邮件调查回复率的方法。首先,调查邮件封面上字母的内容容易造成差异。应该避免直接说,这是一项调查或一张问卷,或者恳求帮助。其次,告知对方你的联系机构、回信日期、姓名和地址、"为什么参与调查如此重要"的声明、保密的承诺、研究的用途、参与的报偿、提出问题和评论的地方、感谢的声明、研究者的签名和职称等。调查邮件最好采用私人信件的包装,象征性的报偿越小越好,可以是一支笔,被试在填写完之后可以留下来继续使用。Pressley & Tullar (1977) 认为,报偿的大小将增加回复率。另一种有效的调查策略是,邮寄调查信件之前,电话联系被试,礼貌地要求他们填写收到的调查。最后,随访信件也会增加回复率。有些研究者在初次邮寄之后的一个星期寄出明信片,几个星期之后,再寄出随访信和调查信件。

即使你付出了上述所有努力，也不可能达到100%的回复率。邮件调查最好的回复率是80%左右，而60%的回复率是比较现实的。如果达到这些回复率，研究者就可以估计所获样本的代表性，并比较样本和总体。例如，获得总体在特定人口统计学变量上的分布，如性别、收入水平和教育等。要求被试在调查表上填写这些信息的理由是，为了与总体进行对比。如果在这些人口统计学变量上，样本和总体的偏差较大，这将是一个危险信号，它提示所获样本并不能代表总体，因此将结果进行一般化是危险的。

收集调查数据的另一种方法是电话调查，打电话给被试，要求被试回答一系列标准问题。首先，电话调查比邮件调查更经济、更容易。其次，电话通常比写信更加私密，取得被试的信任更多，获得的回答更全面。但是，这一方法也存在诸多问题。首先，通过电话得到一些人的回复通常是困难的。白天，很多家中没人；晚上，很多家庭要应付一些推销产品和服务的电话推销员，这些电话推销员经常以做一项调查为开场白，来展示自己推销商品的口才。大多数人会挂掉电话或者礼貌地拒绝，这是可以理解的。越来越多的人将这些电话推销员列入来电被拒绝的名单之中。现在很多人有两到三部电话，有些人还有未使用的号码。即使你得到一个被试的回应，但通过电话让回答者完全理解问题也比较困难，特别是在开着电视或者婴儿啼哭的情况下。这时，问题必须简短，也必须限定答案选项的数目，避免超过回答者的注意广度。对于一些十分敏感的话题，回答者并不确信是否可以保证隐私；他们也无法确定，你就是你所说的那种人；他们还担心，电话旁可能有其他人会偶然听到谈话的内容。

有时为了节省时间和精力，我们可以使用群体调查。群体调查是调查者访问一群被试，在短时间内分发和收集一个书面的调查。例如，上一学期，我要求基础心理学班级的学生回答一页纸的调查，问题是对核电站的态度。分发、填写和收集调查的总时间大约是10分钟。我所在的大学要求上课时间必须充分用于教学，因此，分析结果之后，我花费第二节课的部分时间向学生们呈现了调查结果，同时以此为例解释了心理学研究方法。本班级的许多学生也经常在班级实施调查、收集数据，他们要求40名或者更多的学生来到班级填写问卷。当然，学生们是自愿参加的，也可以不参加。不参加的学生也能通过其他方式完成课程要求。群体调查的优点是，它能够快速提供许多数据。缺点在于难以保证隐私，因为回答者通常坐得十分靠近；也无法得到一个有代表性的样本，因为群体是预先建立的。你认为基础心理学班的学生们能代表一般的群体吗？这些群体通常是被挑选或自我挑选的，所以，并不是感兴趣群体的一个代表性样本。

面对面访谈是一种耗时却有效的调查技术。被访者与采访人面对面坐在一个地方，如研究室、被访者的家或工作室。在结构性访谈中，采访人必须根据事先准备的讲稿提问；而在非结构性访谈中，采访人可以自由发问，与被访者自由讨论一些话题。结构性访谈的控制性更强，数据分析也较为容易；而非结构性访谈对于被访者而言通常比较自然，也可能引发更深层次、更细节的回答。综合考虑，研究者通常前一部分进行结构性访谈，后一部分进行非结构性访谈。

在考虑调查技术时，我们不应该忽略网络这一工具。网络调查完全是电子时代的一种产物，被访者可以直接在网上作答。研究者可以使用网络联系被试，直接将

调查传送给对方，具有高效性。例如，我和一位同事刚刚完成了一项网络调查。我们属于一个特定的职业协会，想从一个代表性样本中收集一些调查信息。首先，我们从机构成员名单中随机挑选样本，通过 e-mail 将一页纸的调查发送给他们，被访者下载填写后再传回给我们。对于一些没有 e-mail 地址或者不工作的人来说，我们进行邮件调查。据我所知，目前，还没有哪个研究比较过网络调查和邮件调查的回复率；但我猜测，对于简短的调查而言，网络回复率通常更高，因为 e-mail 的随访也很容易传送。未来的研究应该制定一些网络调查的规则。

另一位同事还通过建立网页收集调查数据。他要求参与者在网上评定电脑制作的脸部图片的漂亮等级。然后，分析判断漂亮与哪些面部特征相关。他定时升级网站，所以，参与者在回答调查之后，能够比较自己与其他人的等级评定结果。结果的反馈吸引了很多人参与，得到的回复有数千份。因为网络调查的回复者都是自己选择的，所以样本的代表性存在明显问题。但是，对于有些研究而言，这并不是一个大问题。网络调查的被试对于面部漂亮的评定标准与其他人的标准存在差异吗？可能有，也可能没有。不管怎样，我们都应该记住，网络调查通常会得到一个存在偏差的样本。因为并不是每个人都能上网，不同阶层的群体使用网络的比率也有很大差异。因此，你应该了解很多建立网站的知识和这方面的研究者。同时，利用网络中的教程来帮助自己设计调查。如果你认为自己的调查研究可以使用网络，你应该通过浏览器或者寻求教授的帮助，以发现这样的网站。

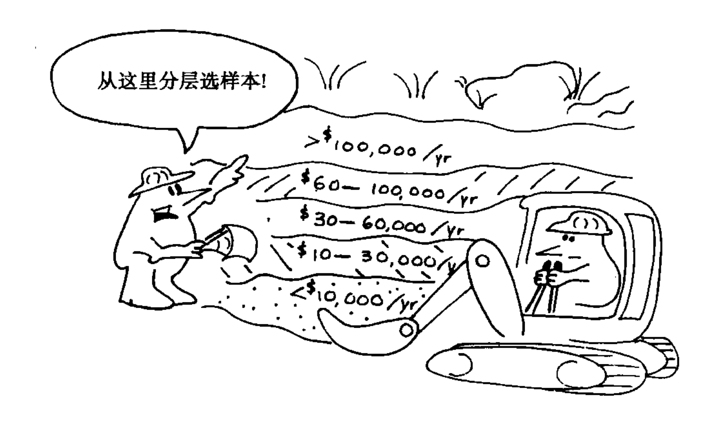
选择样本

我曾经一再强调，研究者在选择调查样本时，希望得到一个具有代表性的群体，并将该群体的结论推广到总体。那么，如何挑选一个代表性样本？一种方法是随机抽样，这一方法通常比想象的要困难。例如，你想将调查结果推广到整个美国。首先，应该列出所有美国公民的名单，然后从名单中随机挑选样本。如果将所有名字放入一个帽子中，这将需要一个巨大的帽子。大多数情况下，你将会从一个较小的总体中挑选样本，如一座城市、一所大学、一个班级等，并期望样本与你实际感兴趣的总体没有太大差异。但是，随机抽取经常不可能实现。大多数心理学系的被试来自于选修心理学引论的学生，这些学生或是自愿完成课程要求，或是为了获得额外学分。有些院系还付给被试报酬。在这两种情况下，被试都不是随机选择的。我相信，你能想到许多抽样方法，但抽取的样本与真实的随机样本完全不同。请看一个非随机抽样的例子。1993 年 Ross Perot 要求电视观众回答 TV 导刊明信片上的一系列问题。调查结果表明，97% 的民众支持大幅度缩减政府开支。但是，随机样本对同样问题的回答却显示，只有 67% 的人支持缩减（Tanur，1994）。

在随机抽样不可能实现的情况下，有些研究者会使用分层抽样，即根据一定的标准将整个总体分成不同的层①；然后，从每一层随机抽取被试。例如，总体是整个美国的公民，研究者可以按照收入水平进行分层：低于 20 000 美元；20 001 到 40 000 美元；40 001 到 60 000 美元；60 000 美元以上。然后，按照适当比例抽取每一层的被

① strata 的日常含义与 layers 相似。例如，我们可以从一个山坡的侧面看见岩石的层级。通常，研究者根据收入水平分层是比较容易的。

试。如果总体是一所大学的学生，研究者还可以根据性别、民族、社会等级来分层。但是，必须确保按照适当比例抽取每一层的被试。



编制调查问卷

完成抽样之后，研究者面临的任务是，如何编制调查问卷？首先，你应该确定是否需要编制这份问卷。如果问卷调查的目的是判断人们特定人格维度，如权威的、焦虑的、内向的、富有创造力的等，或其他人格特征。这份问卷很可能已经有人设计好并出版，而且具有效度和信度（参见第 7 章）。你可能在检索文献时能够发现它（参见第 6 章），也可能在导师的帮助下找到它。因此，你的任务是确定这份问卷是否适合你的调查研究。但是，如果问卷作为杂志文章的一部分已经出版，你将受到版权的限制。如果问卷来自商业出版机构，你会买到它，但价格却十分昂贵。必须坚决抵制从问卷上抄袭问题的诱惑。这一行为不仅是不道德的，而且关系到版权问题，也是违法的。

如果你感兴趣的问题十分具体，或者没有找到合适的问卷，你必须自己编制问卷。这看似简单，你所做的只是提出所有问题，是这样吗？例如，考察人们对流产的看法，你可能会直接问：你如何看待流产？这是一种开放式问题，因为你已经允许被试用自己喜欢的方式来回答。设想被试的回答可能是几句话、几段话，或是一本书，被试回答问题的态度自始至终都十分认真，但是面对这些数据，你如何分析呢？因此，编制问卷的第一步是，以访谈的形式获得人们对问题的可能答案。然后，根据人们的回答编制更加限制性的、封闭式的问题。在整个过程中，你必须选择一个时段来分析收集的数据。理想的情况是，能够将这些数据转换成数字，进行定量分析。避免遇到统计学家经常提到的一种糟糕的情况，即一些学生抱着一大堆问卷走进办公室，将厚厚的问卷放在桌子上说：问卷都在这儿了，现在我该怎样分析？

对开放式问题的答案进行数字化分析也不是不可能的，尽管这需要付出很大努力。例如，一些研究者训练法官独立地阅读开放式问题的答案，然后将这些答案编码成预先设定的答案选项。通过比较每位法官的答案，来判断编码方案是否具有信度。其中的关键之处在于，无论使用何种方法，完成问卷编制的时候，你应该准确地知道你将得到何种类型的数据以及如何分析这些数据。

将被试的答案转化成数字的一种方法是使用封闭式、多选项的问题，这种做法可以限制被试给出其他可能的答案。例如：

在什么情况下，可以允许女人流产？

——从来不；

——强暴或者乱伦的情况下；

——强暴或者乱伦的情况下，且经过父母同意，但父母的同意处于次要位置；

——只要她决定这么做，无论何时都可以。

指导语将提示被试只能选择一个答案，并标上记号。通过计算每一项答案上作记号的回复者的数目，实现数据的数字化。

尽管此类问题可以提供量化的数据，但也经常受到批评。例如，有些回复者可能会思考单词"allowed"的意思：谁执行了这个允许？政府，还是上帝？还有些回复者也可能无法准确作答，因为答案选项并不能准确描述自己的感受。另一些回复者可能认为，胎儿的父亲拥有一些权利，或者胎儿的年龄具有重要作用。另外，表达问题的方式也会产生差异。举例来说，面对同一个问题，人们倾向于特定的回答方式，这也会产生差异。

"女人终止妊娠的权利与处理其他的健康问题相似吗？"

"在什么情况下，政府应该限制妇女流产的权利？"

"在什么情况下，可以允许一位母亲夺走她未出世孩子的生命？"

有些回答表明了某种特定的信仰和相关的情绪。一般来说，人们认为，民众的权利应该受到保护；政府的限制应该降到最小；妇女是独立的，但母亲肩负着责任；胎儿虽未成人，但未出世的孩子也是人；终止妊娠虽不是杀戮，却是夺走了他人的生命。虽然，大多数问题并不像例子中的偏向那样明显，但即使我们谨慎避免，也仍然会出现一些微小的偏向。最近，我正编制一份问卷，考察我任职的大学中一门统计课是否适合心理系的学生？首先，我必须承认，我认为这门课并不好教。在我停下来修改问卷之前，我对毕业班的学生进行了问卷调查，其中包含了一个问题：如果你认为这门课程教得不好，以下原因中你会选择哪一项？然后，我列出这门课程存在的诸多可能的问题，不包含任何积极方面。但是，当一位同事指出问题隐含着明显的否定倾向时，我被难住了。

语言上十分细微的变化将导致观点上巨大差异。例如，一项电话调查发现，53%的人认为政府应该在"福利"上投入更多；而23%的人认为，政府应该在"援助贫困"上投入更多①。在前面提到的一项十分相似的调查中，Ross Perot 提出，每增加一美元税收，用作财政赤字和债务缩减的储备金就会减少两美元，你相信吗？67%的随机样本给予了肯定回答。但是，当问题被改写成"你是支持还是反对这项议案：每增加一美元税收，用作财政赤字的储备金减少两美元，即使这意味着像医疗和教育这样的国民计划经费的缩减？"只有33%的人支持（Tanur，1994）。由此可见，细微的语言变化会使人们的回答产生多么巨大的差异。

首先，对于样本个体来说，提出问题的所有语言必须是可以理解的。大多数大学生和教师的交往人群通常具备一些大学教育经历，但是普通民众的阅读技巧和词汇理解并不是很好，对调查样本采用适当的词汇，这是研究者在提出问题时必须牢

① 来自600位成年美国人的电话调查，刊登于1994年5月18—19的 *Time* 和 *CNN*。

记的。其次,清晰表述问题,不至于产生混淆。例如,避免以否定语气提出问题:"女人不应该有流产的权利吗?"检验问题是否可以理解的最好方法是,先将问卷的草稿发给小样本群体,并要求他们评论,小样本群体必须与研究中的大样本群体大体相似。

将答案转化成数字以便作答的另一种方法是使用等级量表。等级量表的不同等级代表了答案的不同等级。例如:Senator Jones 表达了她关于流产问题的观点,你认为她表达得如何?

非常糟糕 1 2 3 4 5 6 7 8 9 10 非常好

一种方法是要求答题者在某个数字上画圈、打勾、画×。另一种方法是画斜线将整条线继续细分。此例已经在量表的末尾提供了言语标签,即参照;因为它们可以解释连续体的意义。第三种方法是,给出数字或斜线画上标记,如不好、相当不好、中立、相当好以及好等。等级的数字也可以变化,典型的情况是从5到10。五点量表通常是最小的量表,主要是因为有些人回避极端。如果五点量表变成三点量表,观点之间差异的分析空间将会很小。

如果你的目的是考察被试对于众多议题的态度,李克特量表将是比较好的选择。该量表给被试呈现一系列表述,询问被试是否同意每项表述。举例如下:

如果一名女子遭强暴而怀孕,流产是合法的吗?

非常同意	同意	中立	反对	非常反对
1	2	3	4	5

指导语要求被试在最能表达自己观点的数字上画圈。如果不使用数字,也可以使用水平线、×或其他标记。对于这一例子,必须测量从线的末端到标记的距离,以便将答案转化成数字。李克特量表的一个优点是,所有问题的答案的量表等级都相同。另一个优点是它的实际效用,即页面左边通常列出各项表述,右边列出量表数字。参照或描述打印在页面的顶端。这种格式节省空间、易于理解。参见图10-5。

关于流产问题的观点

请在右边你最同意的观点数字上画圈。

	强烈反对	反对	中立	同意	非常同意
1. 流产是一种罪孽。	1	2	3	4	5
2. 政府应该补贴流产的贫穷妇女。	1	2	3	4	5
3. 未满18岁的女子不经父母同意流产应视为违法。	1	2	3	4	5
4. 流产手术的医生应该劝告手术者采取其他选择,如将孩子送人收养。	1	2	3	4	5
5. 怀孕妇女完全拥有流产的选择权。	1	2	3	4	5
6. "Day-after"流产药片是合法的。	1	2	3	4	5
7. 经营流产诊所的人应该被拘捕。	1	2	3	4	5
8. 仅在强暴和乱伦怀孕的情况下,流产才是可行的。	1	2	3	4	5

图10-5 使用李克特量表进行观点调查的例子

许多研究者也在调查中收集人口统计学信息,如被试的性别、年龄、民族、教育水平、收入水平、社会阶层、宗教信仰等信息。具体收集哪些信息取决于调查的目的。例如,你的研究目的是考察人们对流产的观点是否受到宗教信仰的影响。问卷必须增加一个有关宗教信仰的问题;然后,根据这一信息对问卷进行分类。有一个

问题需要一再重申，即你在设计问卷时就应该决定如何分析数据，这一点十分重要。关于这些人口统计学信息应该放置在问卷的什么位置，研究者还存在争议。最明显的位置是问卷的顶端，但增加这些项目可能会使被试感觉厌烦，从而无法完成问卷(Dillman,1978)。

优 点

我们已经讨论了调查和问卷的许多优点。调查和问卷提供了一种测评人们的观点、态度、动机和未来行为的方法。同时，也提供了一种快速、便宜地收集大批数据的方法。

缺 点

我们已经讨论了调查研究的一些缺点。例如，收集的数据如此庞大，使得分析变得非常困难，特别是在设计调查时并没有计划如何分析数据时，尤其如此。即便已经设计了数据分析，也需要十分高级的统计技术分析大批量的数据。如果回收率很低，存在未回复偏差，那么，样本的结论将很难推广到较大的群体中。

第三个缺点是，调查实际上是一些相关性的观察，并非实验，研究者并没有操纵自变量。调查的数据是多重相关测量(multiple dependent measures)。因此，我们必须避免对结果作因果性的表述。例如，我们考察人们关于流产的观点与宗教信仰的联系。我们很可能发现，被试的宗教信仰越根深蒂固，越反对流产。如果我们得出结论：宗教信仰引起人们对流产的否定态度，那么，这是十分冒险的。我们只能说，二者相关。我希望大家记住第1章和第2章所讨论的内容，我们必须谨慎解释相关性的数据。

调查和问卷的最后一个先天的弱点是，它们不能直接测量行为，而是自我报告。被试告诉我们的是他们想告诉的，而我们无法证实这些信息。他们为什么会对我们说谎呢？实际上，他们的回答并不真实的原因有以下几点。首先，他们为了保护自己。尽管研究者已经告诉他们调查是匿名性的，不需要在表格上填写姓名，但他们可能认为，研究者会编码表格，以便识别。或者当他们在房间填写表格时，其他人会看见他们的答案。即使他们相信信息是保密的，被试也会考虑信息的用途；也可能认为，研究者会因为一些原因歪曲他们的答案。例如，一个吸食大麻并认为自己的行为是合法的人，可能会违心地回答：她从未在毒品影响下经历过糟糕的时期。她可能意识到，如果许多吸食者报告了糟糕的经历，这一信息将会有助于认定"吸食大麻是非法的"。被试会说谎，甚至是因为他们想引起大家对所属群体或职业的注意。例如，如果一名学生认为，应该采取一些措施防止中学的枪击事件。他就可能报告，他在校园里看到的枪支比实际看到的多。

其次，在某些情况下，被试并不是有意撒谎，而是自欺欺人，同时也欺骗了研究者。受特定情绪的影响，被试可能不愿意承认自己的感受和态度，特别是当自己的感受和态度与社会上广泛接受的规范存在差异时。例如，一名被试强烈否认自己是种族主义者，但同时他所参与的行为却清晰地表明，他是一名种族主义者。因为大多数社会成员并不接受种族主义，人们很难承认这一事实，即使实际情况确实是这样。

因此,在分析调查数据时,我们应该时刻意识到,最终的数据是自我报告的。在讨论调查结果时,我们还应该记住,应该在一定限度内描述结果。我们不应该说,27%的中学生吸食大麻,而应该说,我们知道的是27%的学生报告说吸食大麻。我的祖母曾经告诉我一句话:说和做是两件完全不同的事情。

表10-3总结了这章涉及的各种研究技术的优缺点。

表10-3 准实验、基线实验和调查的优缺点

设计	优点	缺点
准实验	无法进行真实验时,可以采用这种研究。 可以评定内部效度的影响因素。	威胁内部效度的因素仍可能存在。 比传统实验设计更为复杂。 统计分析困难。
基线实验	单个被试提供的结果容易解释,无需统计。 自变量操作的大小和间隔十分灵活。 可以研究发生概率很小的事件。	很难满足其假设(如可逆性)。 无法考察微小却重要的效应。 推广性受限制。
调查	可以考察内部事件(如态度)。 可以快速收集大量数据。	大量数据有时很难分析。 低回收率可能引起未回复偏差。 结果是相关关系,不能推导出因果结论。 自我报告可能是不真实的。

小结

在应用心理学领域,被试的随机分配经常难以实现。在这种情况下,我们有时采用非传统的设计,即准实验设计。非实验设计的结果很难解释,因为存在许多威胁内部效度的因素。这些设计包括单组后测设计,即只测量单个组在接受实验处理后的行为;不对等组后测设计,即第二组被试的选择方式不同,虽不接受实验处理,也会被测量;单组前—后测设计,即测量单个组在实验处理前后的行为。

准实验设计可以消除或估计许多威胁内部效度的因素。第一种类型是不对等控制组前—后测设计。其中一组在实验处理前后被测量,另一组以不同的方式挑选被试,不接受实验处理,但在相同的时间被测量。这一基本设计的变式包括近似前测,即在前测不可能实现的情况下,测量一个与后测相关的变量;分离的前测和后测样本,将不对等组细分,一半在实验处理前接受测量,一半在实验处理后接受测量;多时间间隔前测观察,每一组在接受实验处理前被多次测量。

准实验设计的第二种类型是间断时间序列设计。其中一个组在接受实验处理前后被测量多次。这一设计的变式包括增加不对等无处理控制组的时间序列设计,其中,第二个不对等组以不同的方式被选择,在相同的时间被测量,但不施加实验处理;去除处理的间断时间序列设计,在去掉实验处理之后,进行第三个系列的测量;转换重复的时间序列设计,以不同方式选择各组,在许多时间点对它们进行测量,但在不同时间点施加实验处理。

第二种非传统设计是基线实验,它仅使用单个被试的数据分析实验效应。基线实验经常可以用以评价治疗或医疗干预的效果。在建立一个稳定状态的反应率

(即基线)之后,研究者开始实验操作;直到反应速度达到一个新的稳定的转换状态;然后,研究者通过恢复到原始基线来论证可逆性。基线设计的一个优点在于,它提供了一种分析个体行为重要变化的有效工具。研究者也可以灵活决定何时施加自变量的一个水平以及施加哪个水平。结果也容易解释。但是,基线实验的一些假设(如可逆性)在许多心理学领域无法满足,因此,很难发现微小却重要的效应,以至于实验结果很难推广到较大的群体。

第三种非传统设计是调查或问卷,它通常用于测评一个样本的观点、态度、动机和未来的行为。邮件调查成本较低,样本人群分布广泛。但是,低回收率会产生未回复偏差,即样本的特定部分不成比例的缺失。因此,研究者不能将研究的结论应用于较大的总体。提高回收率的方法有:写上适当的封面字母、小的报酬、提前的接触、随访信件。电话调查比较迅速,更加私密。但未回应仍然是一个问题。另外,也很难得到一个有代表性的样本。群体调查十分有效,但是,被调查的群组往往不能代表整体。面对面访谈的效性(efficient)不如其他技术,却是一种更加私人化的方式,可以采取结构式和非结构式两种访谈。新近发展的网络调查通过网络请求参与,可以通过电子方式或邮寄方式进行回复。

调查样本通常是一个随机样本,总体的所有被试被选中的概率相等;分层抽样是在分层之后,按比例随机抽样。调查问卷可以是开放式问题,它很难被量化;也可以是封闭式问题,它限定了答案选项,如多项选择题或等级量表问题(李克特量表)。尽管调查可以快速收集大量关于观点、态度、行为的数据,但是,较大量的数据分析仍然很困难,低回收率会引起未回复偏差,因果关系也不能通过相关数据推导出来,自我报告可能是不真实的。

关于准实验统计的推荐书目:

对于初学的学生:

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally.

对于高年级的学生:

Box, G. E. P., & Jenkins, G. M. (1976). Time-series analysis: Forecasting and control. San Francisco: Holden-Day.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental design for research. Chicago: Rand McNally.

Kidder, L. H., & Judd, C. M. (1986). Research methods in social relation (5 th ed). New York: Holt, Rinehart & Winston.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

关于基线设计的推荐书目:

Hersen, M., & Barlow, D. H. (1984). Single-case experimental designs: strategies for studying behavior change. New York: Pergamon Press.

Robinson, P. W., & Foster, D. F. (1979). Experimental psychology: A small-N approach. New York: Harper & Row.

Todman, J. B., & Dugard, P. (2001). Single-case and small-N experimental designs: A practical guide to randomization tests. Mahwah, NJ: Lawrence Erlbaum.

关于调查设计的推荐书目:

Dillman, D. A. (1978). Mail and telephone surveys: the total design method. New York: Wiley.

11

如何判断你已经做好了准备

多次失误即成错误。

佚名

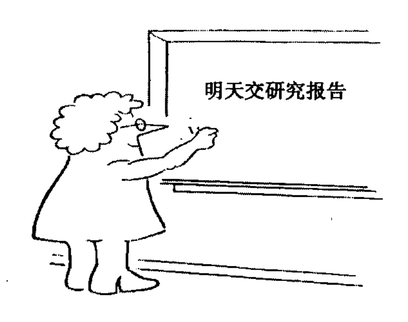
我认为,最大的错误就是不知不觉。

T. CARLYLE (1888)

傻瓜的做法是自以为是,而智者则善于听取意见。

PROVERBS 12:15

你现在应该掌握了开始进行实验的所有工具。但是，在开始收集数据之前，你或许还需要问自己一些问题，以判断你是否已经考虑到了所有重要的事情。



如何知道你可以开始做实验了。

当我教授实验方法时，我的学生必须思考、设计并实施一个原创的实验。在允许他们收集数据前，他们必须向全班展示自己的实验设计。全班同学和我本人的共同工作是，评价这个实验、找出它的缺陷以及考察这个学生是否考虑到了所有重要的细节。这种练习有多重目的。这能够帮助倾听实验展示的学生锻炼一种批判性思维——这是所有科学家必须具有的能力。准备实验展示的过程也能够督促学生思考他们作出的所有实验假设以及到这个阶段为止曾忽略的细节。最后，这种练习给我们大家一个机会来帮助实验者改进实验设计。

好好社会

在讨论那些准备开始实验需要考虑的问题之前，我想首先评论一下我的很多学生对展示实验设计表现出来的情绪反应。展示实验设计的同学把这个过程看成是这门课最令人厌恶和最有压力的部分。一部分压力只是来源于做报告本身，无论是任何报告——这个技能在其他课程上是很少涉及的。但是我怀疑更多的压力是来源于他们必须在可能比较有批判性的群体面前为自己的实验设计辩护。



可能提出批评意见的一些同行。

学生的第一反应是保持沉默："如果你不找我的麻烦，我就不会找你的麻烦。"

即使我鼓励他们,有些学生也不愿意对其他学生的想法进行批判。我们生活在一个"好好社会"里,这里的规则就是对他人的行为和观点表达最大限度的忍耐。一些人似乎相信,因为我们都有权利表达我们自己的看法,一种看法的价值和另一种看法的价值是等价的。批判他人的看法会被看成是对他们自身或对他们言论自由权利的攻击。

从学生在课程结束时填写的问卷来看,确实有一些学生把我对学生报告的评论看成是人身攻击和毫无理由的。虽然我总是尽可能保持微笑和压低声音,并且表现出帮助他们的态度,这些学生看来不能理解为什么他们和蔼的、友好的老师突然将矛头指向他们。

我希望本书的以上各章至少能够在知识水平上说服你,在科学领域,一种观点并不总是与另一种观点一样好。观点必须是可辩护的。如果科学的原则被违背了,那么,结果也就是令人怀疑的和毫无用处的。第 3 章中讨论的演绎和推理的逻辑,就是讨论特定的结果支持或否定某理论的基础。在第 2 章中讨论的消除潜在的混淆变量,则是建立因果关系的必要条件。也就是说,要说明因变量的变化是由于自变量造成的。第 2 章中还讨论过,我们要随机地选择实验被试样本,这是实验结果能否推广到更大人群的基础。

当班级同学、导师或者同事评价某一实验设计时,他们是在试图根据科学原则来帮助实验的设计者。因此,当实验完成后,实验结果才是可辩护的并且可以加入到科学知识体系当中。虽然在实验设计提出阶段的批评可能会令人烦躁,但在研究完成后再提出则是致命性的。事后的批评不仅说明,因为你没有考虑周全而犯下了错误,更说明你浪费了你自己和被试的时间以及其他本可以用来增长知识的其他资源。这个说法并不是说科学是非常严肃的事业,犯错就必须忏悔;而是说科学具有特定的规则,作为科学家就必须遵守。你应该使用任何可用的资源,包括别人的建议,来帮助你遵循科学原则和做出优秀的研究。

在你开始之前的问题

在我的学生进行实验设计的时候,以下问题是我最经常问他们的。你可能已经回答了这些问题。真棒!如果没有,那么,你应当确认你能在收集数据之前能够回答它们。

实验是否符合伦理规则

就像我们在第 4 章中看到的那样,人道地对待实验被试引发了一系列问题。你考虑到了所有这些问题吗?你的被试是否会承受任何心理的或身体的压力?如果是的话,如何能将压力减到最低?你的被试是否填写了知情同意书?如果你使用自助注册表格,这些表格是否适当地描述了你的实验以保证被试能给出知情同意?你能将这些知情同意备案吗?你是否在实验中使用欺骗?如果是的话,你是否在实验结束后对被试进行了适当的解释?你准备了解释的声明吗?你是否准备了时间表,这样你可以按时地会见所有被试?万一设备出现故障或者你生病了,你是否考虑好如何应对?你知道如何保证你收集的数据的保密性吗?你的实验是否含有暗示性

指令？这个问题是否可能影响实验结果？你应该在实验开始前询问自己上述问题。

另外，你要准备好所有的纸质文件来帮助你获得研究机构审查委员会对该研究的许可。在某些情况下，审查委员会需要几个星期来考察你的实验设计，在他们最终同意之前你不能开始你的研究。所以，一旦你确定了实验设计，就快速地填写并提交相应的文件。假如你的实验在伦理上存在争议，如使用欺骗或者存在潜在的压力、涉及药品使用等，那么，遵循这个原则就尤为重要。在这种情况下，审查委员会很可能需要更长时间，他们可能会找你来确定某些细节，也有可能不同意你的研究。

需要多少被试

一般很难估计出你的研究需要多少被试。学生经常犯的错误是使用被试过少，这样看起来很好的结果可能在统计上并不显著。当然，也可能存在一些实际的限制，比如对特定被试的限制导致了你只能使用少数的人。如果是这样的话，你可能需要作出让步。假设你想要尽可能多的被试，那么，有统计方法能够确定统计的功效以及你需要的被试数目。我们会在下一章简短地介绍统计功效，虽然做这些计算实际上不是本书应该涵盖的范围。

你应该时刻牢记，只使用少量的被试将导致原有的实验效应在统计上不显著，而使用过量的被试将导致不重要的效应在统计上显著。在这种情况下，使用过多的被试不仅无效，还可能具有误导性。

决定大概使用多少被试的最好办法就是研读文献。当然，如果你是要重复别人的研究，其中报告了某个显著的效应，那么，你很容易知道该使用的被试数目。即使你不是进行直接的重复，你也可以找到类似的文献，该文献也使用了你提出的因变量。特定因变量（例如反应时，或者一串单词中可以被提取的单词量）产生的数据的变异性是相对稳定的。如果你确实找不到类似实验，那么，你可能需要做一个预实验以测试一下你需要的被试数。

实验是以单个被试做还是以小组集体做

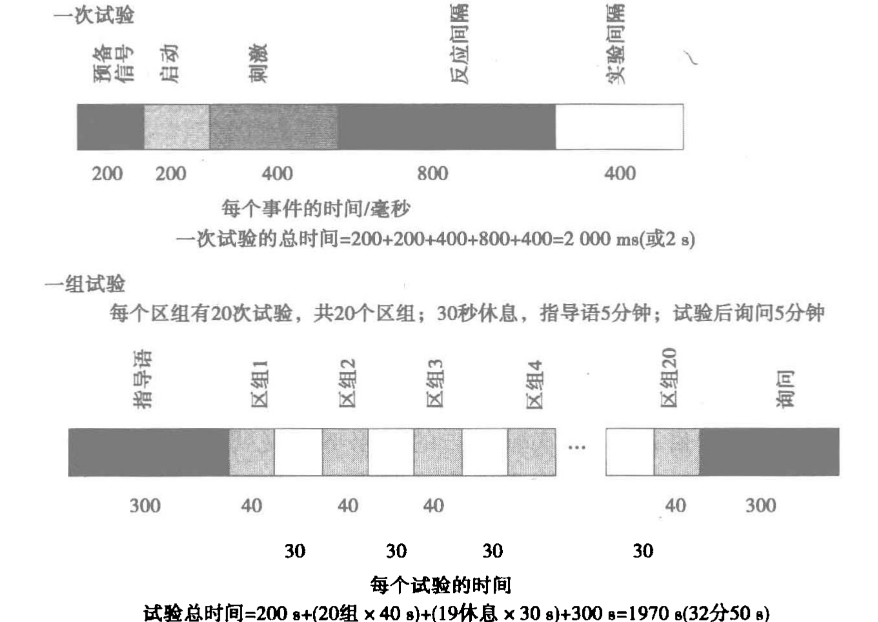
多数新手研究者首先想到的都是让被试逐个来而不是集体来。有时候我们无法选择。例如，如果只有一个设备能够用来记录被试反应；另一些情况下，在组里个别人可能会影响他人的操作，这样就必须单独做。但是，如果你可以让被试成组做实验，那么，就可以更有效率地收集数据。在考虑选择时，你应该自问以下问题：我能否给成组的被试做问卷还是单独问他们问题？我能否用幻灯片或投影仪来给成组的被试呈现刺激材料而不是使用卡片或计算机给单个被试？如果你给被试呈现一系列固定的刺激，并且只需记录一些简单反应或特定类别的反应正确率，那么，是有可能让被试成组完成实验的。如果刺激材料的顺序取决于前一个试验的反应或需要记录反应的精确时间，那么，就需要让被试单个来做。

实验需要多长时间

应该从几个层次上考虑实验所需的时间问题。在最宏观的水平上，收集数据需要多少小时，或者多少天，或者多少星期？在最精确的水平上，单独的一个实验段需要多长时间？如果你要使用单独的试验，你必须知道单个试验需要的时间和需要多

少个试验才能回答上述问题。为了确定试验长度，你必须知道一个试验内发生个各种事件的顺序、各个事件需要的时间以及试验间的间隔（即试验之间的间隔时间）。然后，当你知道实验需要多少次试验时，你就可以确定完成整个实验的时间。在有些情况下，试验的数目可能需要进行估计。例如，当你要求被试必须达到一定的操作标准，如被试需要在连续的两次试验中正确回忆起一个单词列表。

你还应该牢记的是，将与实验相关的其他任务的时间也包括在内。一般情况下，需要给被试呈现指导语，也应该允许他们在开始前提问。如果实验需要被试进行一些学习或者需要相对稳定的操作，你也许还需要安排一段练习。在较长的或者无聊的实验中，需要中间休息来避免疲劳。实验结束后可能还需要一段交流的时间，尤其是作为课程要求学生来参加实验或者他们期望了解实验的内容。最后，还需要一些“额外的”时间，因为有些时候被试会迟到。如果没有这些时间，那么，连续的实验时间会逐渐地越来越晚。图 11-1 显示了计算一个实验段所需时间的各个步骤。



再加上一些迟到的时间、寻找的时间等，这个实验应该安排 40 至 45 分钟

图 11-1　一个计算实验所需时间的示例。在这个实验中，每个试验包含了 200 毫秒的提示信号、200 毫秒的启动信号（刺激呈现前的信息）、400 毫秒的刺激呈现时间、800 毫秒的反应间隔时间以及 400 毫秒的试验间间隔（两次试验之间的间隔）

然后，你要确认这个实验用来收集被试数据的总体时间是多少。同样地，你应该留出一些灵活的时间来采集额外的被试，以填补那些可能缺席的被试，或者那些由于不符合特定标准而被删除的被试，或者由于机器故障而浪费的被试。请记住，对于完成整个实验需要的时间而言，数据收集只是一部分工作。你还需要时间来分析数据、解释数据以及准备实验报告的草稿。这些工作可能都比你预计要花的时间长。

需要限制被试群体吗

一般来讲，是否限制参加实验的被试群体取决于你想将实验结果推广到何种群体。如果你从特定群体中选择被试，这就减小了结果的推广力。例如，如果你使用大学生被试，你必须考虑到有些年龄过大，而有些年龄过小。另外，你还必须考虑到，与一个国家的平均人群水平相比，大学生具有较高的智商、较好的社会经济境况、较高的阅读能力，并且不太可能有健康问题。所以，你不能理所当然地把结果推广到整个人群。

但是，由于实践方面的原因，你可能希望删除个别被试，即使这样做可能会进一步限制结果的推广能力。例如，如果你研究语言能力，诸如阅读、识字、记字或其他可能的语言任务，你可能需要限制被试的母语为英语。如果你研究视知觉，你可能需要使用矫正视力为 20/20 的被试以及可以通过色盲测验的被试。如果你研究运动能力，如运动心理学，你可能需要删除那些有身体缺陷的被试，因为他们的数据很可能没有用。在某些情况下，你可能只能使用女性或男性被试；但在另一些情况下，你需要使用等量的男性和女性被试来考察性别效应对操作的影响。这些例子只是一些你需要考虑的对被试的限制。根据实验中的特定任务，可能也需要考虑其他适当的限制。

应该设定事前的标准来删除被试吗

就像我们在第 4 章中讨论的那样，你可能需要在收集数据前设定操作的标准。例如，我经常在实验中收集反应时数据。经常会有一到两个被试的整体操作水平与其他被试不能相提并论。所以，我经常设定一个标准，如果任何一个被试的平均反应时超过被试整体平均反应时 300 毫秒，这个被试就会从数据处理中被删除。这个标准以及删除被试的数目，需要在实验报告的结果部分说明。就像在第 4 章中提到的那样，不能通过删除被试或数据来支持你的假设。

设定前期标准的总体目的是排除那些与其他被试显著不同的被试，因为这些人会加大数据的变异。他们可能在动机因素、人格因素或其他个人因素上与其他人不同。研究个体差异或变态行为的心理学家可能对这些差异很感兴趣。但是，大多数实验心理学家对这些行为不感兴趣，实验心理学家所关注的是如何建立研究行为标准的科学。

能操作性地定义所有变量吗

在第 7 章我们讨论到，你必须能够对自变量和因变量进行操作性定义，并且详细描述这些操作，以保证其他研究者能够重复你的实验。虽然你应该在设计实验的初期就确定操作性定义，但有些研究者却没有这样做。

因为自变量是研究者最感兴趣的变量，必须非常小心地进行定义。设想你现在需要做一个实验，这个实验是我的学生最常提出的①：你想检验听音乐对学习的作

---

① 这个实验非常类似于测试药物 X（如大麻、酒精或可卡因等）对任务 Y（如驾驶、学习或记忆等）的影响的实验。如果老师要你提出一个原创的实验，那么，请你不要提出类似的实验。

用。这个基本实验的变式可以是检验摇滚乐相对于古典音乐的作用,检验电视的声音、噪音以及重金属音乐相对于轻音乐的作用。假设实验的比较是关于摇滚乐和古典音乐的作用。什么是摇滚乐? 重金属、庞克、新浪潮、嘻哈,还是摇滚? 什么是古典音乐? 施特劳斯的华尔兹、德沃夏克的新世界交响曲、贝多芬的圆舞曲,还是柴可夫斯基的1812序曲? 即使你使用的都是古典音乐,当你使用弦乐四重奏或者使用滚动定音鼓和卡农演奏时,实验结果可能是完全不同的。

与之相似,如果你要比较重音乐和轻音乐,问题是,重音乐要多吵? 一个合适的答案绝不是:"我会把声音开大,直到听起来很吵"或者"我会将音量设置为8"。另一个研究者将不知如何设置来取得同样的音量。理想情况下,你应该请人用仪表来测量这个声音的音量,从而告诉音乐的平均分贝数。

对因变量来讲,操作定义也是必须的。你要怎样测量音乐是否对学习有效果? 这个问题有很多可能性。你可以测量在一定时间内被试可以阅读的页数;你可以测量被试可以解决多少数学题;你也可以给他们做一个测验。你还可以让被试评定在某个条件下他们感受到的学习难度。以上的每个测量都有优点和缺点,但你必须对因变量进行操作化定义,否则,实验设计是无法形成的。

应该事先安排好实验所需的仪器和材料吗

许多实验都需要设备,而几乎所有的实验都需要某类型的设备材料。幸运的是,计算机可以用来精确地呈现刺激、记录和储存被试反应。如果你可以使用这样的仪器、知道怎么使用或者可以得到内行的帮助,那么,你只需要很少的准备就可以开展很多实验。但是,如果没有现成的计算机可用,或者你的实验不方便用计算机来执行,那么,你可能需要用老办法来进行实验,使用手头可用的仪器甚或自己动手来组建仪器和材料。在某些情况下,你甚至需要根据手头可用的资源来计划实验。

你需要准备的材料包括指导语、记录反应数据纸张和事后询问的脚本。你应该提前写好指导语。之后你可能需要在写实验报告的时候将它们作为附录。在任何情况下,只要有人想要重复你的实验,必须能够获得这些指导语。

简单地把指导语拿给被试并期望他们能够认真领会并不是个好主意。有些人的阅读理解能力有待改善,尤其是当主试在旁边徘徊并希望他们赶快读完的时候。研究者应该给出书面的指导语,同时大声地缓慢地读出来。(被试毕竟是头一次听到这样的指导语,虽然主试已经读过很多遍了。)指导语的最后一句通常是"你有没有什么问题?"即使在有些实验中,学习效应不是个问题,但给被试适当的练习是有用的,因为这样他们才知道一旦实验开始将出现什么。这些练习一般安排在指导语之后。

如果你的实验需要记录被试的反应(例如,相对于做问卷来讲),你可能需要为此而准备反应记录纸。如果自变量的不同水平呈现的次数不同,那么,反应记录纸也应该包含这方面信息。如果随机地呈现不同类型试验,你应该使用随机数码表或者其他的随机工具提前确定随机顺序。如果你在实验进行的时候产生随机序,它们常常并不是随机的(见第2章)。

最后,你应该写一个简单的脚本,这样在实验结束后可以告诉被试实验的目的。当被试是为了完成某个课程而参加实验,这种脚本常常是必须的。即使被试不需

要，写个脚本也不多余。被试将在完成实验后感觉更舒服，不会对他们刚刚做的实验产生误解。另外，他们也许可以从这个实验学到一些关于心理学的知识。他们至少会感觉比较好，因为有人花了时间来给他们解释这个研究的目的以及感谢他们的合作。



需要知道怎样分析数据吗

第 12 章将探讨如何使用描述性、推断性的统计学来解释实验结果，附录 A 是一份某些常用统计检验指南。如果你的实验比较简单和短小，这些就足够帮你确定如何分析数据。如果你的实验比较复杂，你可能需要导师、统计书籍或者统计咨询师的帮助，以选择一个合适的方法进行数据分析。无论你如何选择合适的统计方法，都应该在实验开始之前就清楚如何分析数据。统计咨询师经常描述那些找他们的人的恐怖故事，这些人在实验后拿着大量的数据，但最终发现它们毫无用处，因为这些数据无法分析！不要成为这种故事的主角。在实验之前就弄清楚你的数据应该使用何种格式以及怎样分析处理它们。

该怎样解释可能的实验结果

当你决定去做一个实验，你对结果的模式可能已有了一些想法。与统计当中的零假设不同（操纵自变量将不引起任何的效应），你大概期望某些自变量的水平会引起因变量的差别。科学家应当是没有偏见的中立者，而不应当积极地期望某种结果。而事实上，科学的重大乐趣在于预测结果。成为好的预测者是成为好的科学家的重要部分①。但需要小心的是，你不应该沉迷于某种预测而丧失了客观性，从而得出带有偏见的实验结果。

现在请准备好不带偏见地解释你的实验结果。由于其设计的原因，在某些情况下一些实验会被认为是失败的。科学家将这些支持零假设的实验结果称为阴性结果。例如，假设通过一个实验测定摇滚乐和经典音乐对学生学习能力的影响。结果表明，听摇滚乐和听古典音乐的学生的成绩没有统计差异。像在第 12 章中将讨论的那样，因为你的统计检验是为了检验差异，而不是一致性，你不能说这两组的表现是一致的，你只能说这两组的表现没有表现出差异。去研究一下何种摇滚乐和古典

① 英国教授可能称之为矛盾修饰法（oxymoronic）[我希望不是低能的（moronic）]。

音乐对学习的影响是没有差别的,可能也是一个有趣的问题。没有观察到该差别可能除了其本身没有差别外还有很多其他原因。例如,被试量不足、没有控制好变量从而导致数据的变异很大等。在有些情况下,如果一系列使用相似实验条件的实验得到了一致的效应,你也许可以自信地认为,你的阴性结果是有意义的。一般情况下,阴性结果是没有意义的,除非它作为一个方法上的范例说明某些东西不要去做。无论你的结果是阴性结果还是阳性结果,你都应该接受这个结果并试图进行解释。一般情况下,实验者会倾向于不理会结果并简单地归结为设计或方法的原因。但是,无论你在实验开始时多么强烈地相信自己的假设,一旦实验结束,你必须接受该结果并试图进行解释。

决定你是否能够解释实验结果的一个办法是考虑该实验可能出现的任何结果,并考虑你是否能够解释它们。就像在第3章中讨论的那样,一个理论能够帮助你进行预测。请记住,一个理论是对一些抽象的变量间(如果是实验理论,则是关于自变量和因变量之间的)可能关系的陈述。这比针对某个实验的特定结果的描述更加一般化。你需要确认实验是否符合该理论框架,以保证你预测的实验结果和该理论所预测的实验结果相同。如果两个或更多的理论对实验结果作出了不同预测,那么,则更为有利。如果其中一个理论支持阳性结果,而另一个理论支持阴性结果,那么,当可能的结果发生时,你将更容易进行解释。但是,最好的情况是两个理论都预测阳性结果,只是差异的方向不同。在这种情况下,任何的结果都可以清楚地被解释。

第二个进行预测的方法是参考前人的研究。有人可能已经完成了与你的实验有部分相同的实验。在这种情况下,你可能会预测你们发现相似的结果。如果是这样,那么,你可以证明该结果能够被重复,并且它可以被推广到稍有不同的实验情境中,于是,你可以作出更具一般性的理论推断。如果你的结果与前人的结果不同,那么,你可以发现前人结果的局限,同样可以得到一些有用的启示。

第三个进行预测的方法,尤其是在有些领域还没有被研究,尚缺少理论时,可以进行逻辑推断。例如,你可以根据逻辑假设,嘈杂和不可预测的音乐是干扰学习的,因为这种音乐会将学生的注意从学习任务上转移。你还可以使用类似的推断进行逻辑推理。例如,能够将嘈杂和不可预测的声音屏蔽掉的音乐是对学习有帮助的,或者学生对某音乐越熟悉,其对学习的干扰作用越小。虽然这些假设在开始阶段是基于逻辑的,如果得到你的实验支持,则它们也可以上升到理论水平上。

你应该能够预测实验结果的根本原因是,在实验开始之前你就知道自己的实验结果能够对科学提供一些重要的证据。如果你不能说明实验结果支持任何重要的观点,那么,该实验就没有对科学作出贡献。例如,如果你的结果与过去所有理论预测的一样,不能排除任何可能的理论,那么,你的工作对该知识体没有任何新的贡献。像在第3章中讨论过的那样,科学的进步在于证伪理论,而不是证实某个理论,而你却不能证伪任何理论。基本的观点在于,如果你认为,某个结果对于科学的进步意义重大,那么,你应该在做实验前就能够予以说明。否则,它在科学发展中就没有任何意义。

现在准备好了吗

如果你已经回答了上述的所有问题,你大概已经做好了开始实验的准备。最

后，你还应该确认一下，你是否能够将除实验结果外的其他部分都撰写出来。事实上，这样做可以节省你很多时间。实验心理学领域的很多研究生在做论文之前需要提交一个正式的计划书。这个文件基本上类似于最后的实验报告，只是该文件的结果部分是几个可能的预测结果而不是真正的结果。当然，该文件也不包含统计分析。制作实验计划书的好处是，大部分的写作已经提前完成。从研究生院的层面来讲，这个过程能够在某种程度上帮助学生，因为教学委员会可以在实验前指出学生设计可能存在的重大缺陷。对你自己而言，尝试提前撰写实验报告的主要好处在于你必须在写作前回答本章提出的问题。显然，如果你不能整理出实验的所有细节，则无法将它们写下来。没有文献搜索则无法撰写文献综述；不对相关的理论和过去的实验结果有所了解，则无法预测实验结果。撰写实验计划书是个简单的好办法，它可以考察你是否全面地考虑了自己所要做的实验的所有细节。

此时，你应该准备好了，可以进行实验和收集数据。设计一个实验是很有趣的、很好的智力练习。搜索文献需要一些特定的方法，也很有趣。找到合适的统计分析方法会让某些实验者很兴奋。但是，坦白地讲，我进行统计分析只是因为它是实验的一部分。收集数据、检验理论和假设等创造性工作值得你付出努力，尽管这是一些看起来不怎么有趣的过程。通过自己的工作，作出了对知识体系产生深远影响以及有意义的贡献，从而推动科学进步，我确实从中得到了满足。但是对我而言——我希望对你也一样——作为科学家的最大乐趣在于发现以及寻找其他人从未发现的事实。这就是为整个科学事业作贡献的本原力量。

小　结

在你准备好进行实验之前，有很多具体的细节需要考虑。确定你是否预测了这些细节的有效方式就是你，或通过口头介绍，或以书面说明书形式，向其他人提出你的各种实验方案。一项重要事情是你的实验是否满足所有的伦理道德标准以及该实验是否得到伦理委员会批准。在决定需要多少被试时，你可以参考文献中报道的类似实验并使用相似的数目。另外，你需要确定让被试单个测试还是集体测试。为了确定完成实验需要的时间，你需要确定每次试验持续的时间、试验的数目、进行其他活动的时间以及每个条件下的被试数目。在确定你是否对选择被试进行限制时，你应该考虑自己希望将实验结果推广到何种人群。为了避免在实验数据中引入噪音，你可能需要设定剔除被试的标准。为了对变量进行操作性定义，你必须能够清楚地定义操作自变量的步骤以及测量因变量的方法。安排好需要的仪器和材料，它们通常包括指导语、反应记录纸以及事后询问脚本。最后，你应该清楚地知道你将如何对数据进行统计分析以及如何解释结果。现有的理论、过去的发现和逻辑推断都能够帮助你进行结果解释。如果你在实验前考虑了所有这些问题，那么，完成实验和报告结果等工作将会变得更加顺利。

12

如何解释实验结果

包装精美的统计比希特勒的"超级谎言"要强一些；虽然它有点误导，但不会强加于人。

D. HUFF (1954)

谎言有三类：一般谎言、弥天大谎和统计。

BENJAMIN DISRAELI

我相信，多年以来，人们过度依赖于通过假设检验获得的贫乏而又模棱两可的结论，它已经不知不觉地将我们带入概念的死胡同，这个死胡同妨碍了我们的视野并掩盖了我们的潜力。

G. R. LOFTUS (1993)

尽管统计常常被误用，但在面临不确定情形时，它仍然是一个一流而又强大的决策工具。

ROGER E. KIRK (1990)

现在你做好了收集数据的准备。如果你记录被试的手动反应,那么,为每个被试建立反应记录表是很有必要的。这个表格应该包括被试的编号和性别、实验条件以及其他你想要记录的关于这个被试或者关于实验阶段的某些信息。显然,表格里应该留有足够空间来系统地记录被试的反应。实验之后,可以根据自变量和它们的水平对这些信息进行分类。如果你的实验是在电脑上完成的(类似于典型的实验),程序就能够帮你整理数据。如果你是手动收集数据,那么,你需要自己整理数据。如果你需要使用电脑来进行统计分析,那么,你需要在电脑上整理数据。现在,假如你有一系列数值,那么,你要做很多工作才能回答这个有关实验的问题,即自变量水平的变化如何影响因变量?为了回答这个问题,你需要知道分析数据的几种方法以及如何使用它们。

本章将会帮你了解数据分析的潜在逻辑,但不会帮你分析实验的具体统计方法。如果你需要做这样的计算,你应该首先阅读这一章,然后,阅读附录 A 中相应的统计方法。这个附录的本意并不是对统计方法的补充,而是让你能够在老师的帮助下分析本书所讨论的许多简单实验。这一章或附录都不能代替统计课程。我们在这里只讨论基本的统计方法,它能帮助你选择一种统计分析方法,这样你就能够分析简单的实验。如果你需要做更进一步的实验操作,那么,基本的统计课程是必需的。

频率分布

假设你对心理专业学生的焦虑水平是否与经济专业学生的焦虑水平不同感兴趣。你找到了学校的各专业学生名单,从这两个专业中随机各抽取 10 名学生,并且说服他们来参加一个测试。这个测试能够说明一个人的整体焦虑水平。这两个组的测试成绩就是原始数据。①

表 12-1 中是从 0 到 100 之间的虚拟数据。数字越大,学生的焦虑水平越高。那么,这两组之间有差异吗?查看这个案例里的单个分数无异于聆听一首歌的单个音节,你将很难体会整体的节奏。你需要重新安排原始数据使得你能够更容易地解释它们。你可以描绘频率分布图,这就是简单地描绘数据中每个分数出现的频率。但是,你可能已经注意到,没有一个分数出现超过一次。这样,为了使得分布有意义,你需要把单个分数归成类。你希望在更高频率出现的组内有几个数据,那么,你可以在每组内包含 10 个分数(例如,从 10 到 19)。图 12-1 描绘了你的两个实验组的频率分布图。标识为"频率"的纵轴就是落入每个分数组的原始数据的数目。

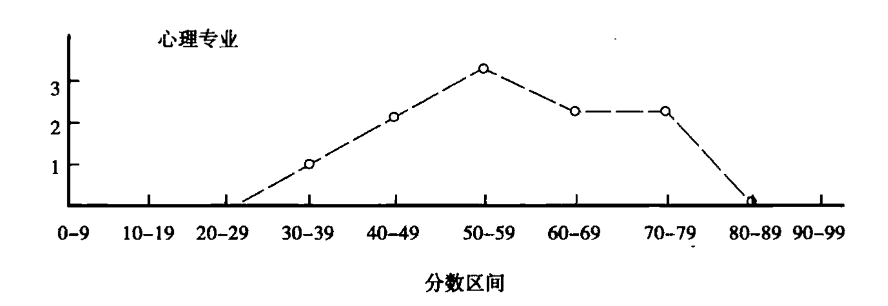
频率分布图是发现条件间差异的第一个有用的方法。有时候实验效应比较明显,对分布的直观观察就能够告诉你条件间存在差异。但是,在这个例子中,两组的分布看起来很相似。

统计学家已经命名了不同类型的分布,研究者可以在对话中使用共同术语,而不必给对方看你的整个分布图。我们已经提到了正态分布(如图 12-2 的左上角)的

① 你可能注意到,这不是一个真正的实验,而是一个相关研究,因为你是在比较两种行为,即选择专业的行为和回答测验中问题的行为。研究中并没有对自变量进行控制。

表 12-1 10 名经济专业和 10 名心理专业学生的焦虑分数

经济专业		心理专业	
学生编号	分 数	学生编号	分 数
1	62	11	55
2	56	12	42
3	67	13	61
4	91	14	58
5	53	15	70
6	87	16	47
7	51	17	62
8	63	18	36
9	46	19	74
10	71	20	51



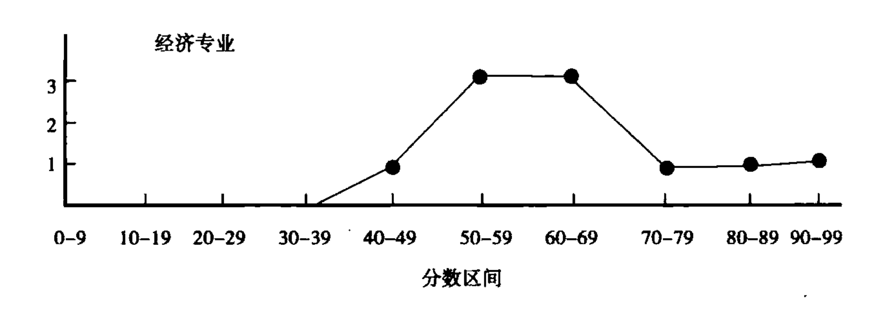
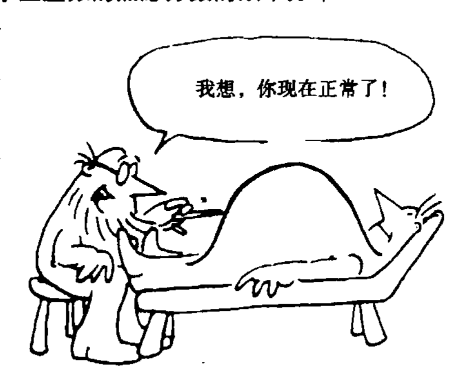


图 12-1 表 12-1 中的经济专业和心理专业学生虚拟的焦虑分数的频率分布

一些特点。一个分布需要符合复杂的数学公式才可以被称为正态分布。但是，对我们来讲，如果一个分布看起来是类似图中所示的钟形分布，那么，我们可以简单地说这个分布接近于正态分布。了解你的实验数据分布是否类似于正态分布是很重要的，因为你想要用的很多统计分析方法要求你的数据接近正态。



其他一些类型的分布也在图 12-2 中可见。具有两个而不是一个高频峰的分布称为“双模态”分布。包含男性和女性的群体身高的分布常常是双模态的。如果一个分布在

某个方向的尾部有更高频的分数，从而导致不对称的分布，则称为偏态分布。博士生的智商分布应该是偏态分布，因为一般来讲，智商很低的人较少。但是，如果一个分布某个方向的尾部像被完全切掉了，则被称为截断分布。反应时的分布应该是截断分布，因为被试反应的时间常常有个速度限制①。

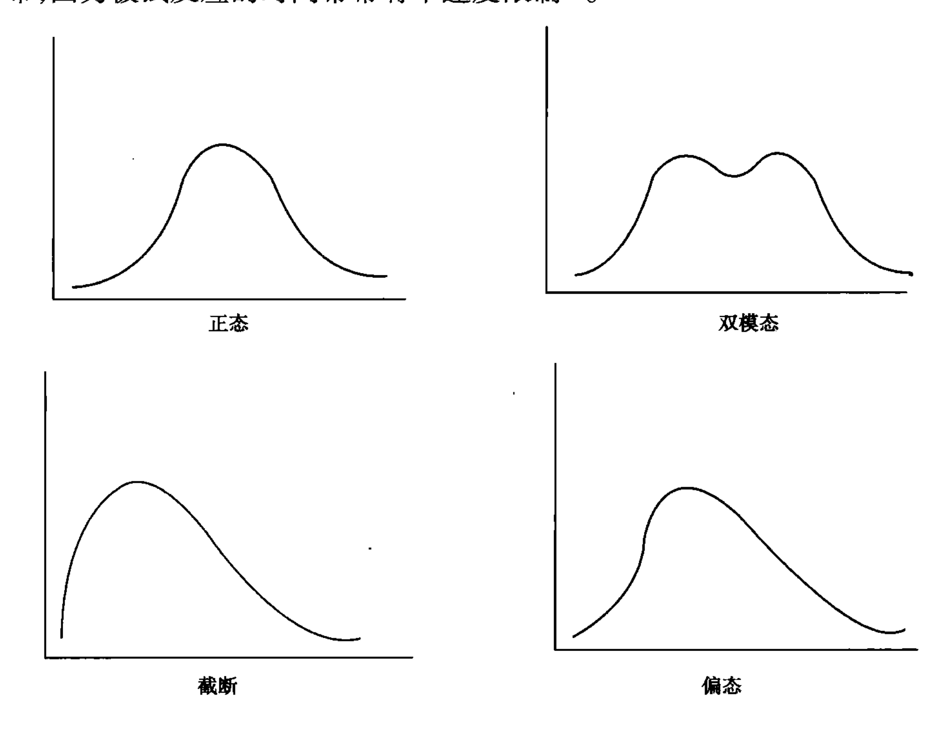
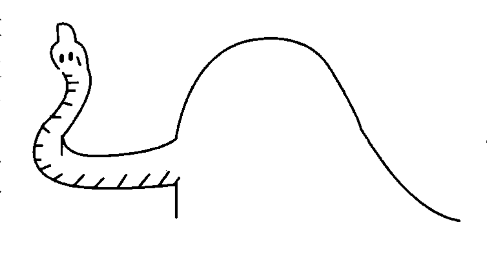


图 12-2　4 种频率分布

描绘频率分布使你能够有序地描述数据，而不是简单地列举出原始数据。但是，这仍然是表述实验数据的一个笨拙方法。能用一个简单数值来描述每个组的被试表现情况则更为方便。我们需要的是计算该描述统计值，并用它来描述数据。



截断分布

描述统计

心理学家通常使用两种基本的统计方法，即描述统计和推论统计。描述统计就是使用简单的数值来帮助实验者描述数据的某个特点，而无需报告每个数据。推论统计将于稍后在本章进行讨论。

集中趋势

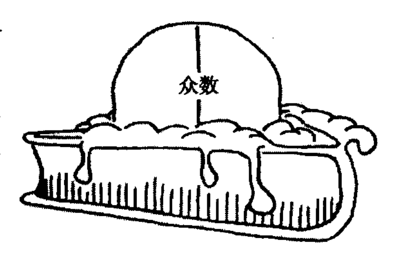
我们想要从一组数据中得到的重要内容就是被试在不同条件下的典型行为。心理学家将这种描述典型行为的统计称为集中趋势。在我们的例子中，对两组进行

① 还记得天花板效应吗？它们常常能够导致截断分布。

比较的一个方法就是计算心理学专业和经济学专业学生的集中趋势。

通常有三种方法来表示集中趋势。众数(mode)是最容易计算的统计值,但这通常也是最坏的一个,因为它忽略了很多数据。简单地说,众数就是出现频率最高的那个值。在我们的例子中没有众数,因为没有一个数值出现次数超过1次。当我们把数据归成组后,可以发现众数。心理学专业的众数是50—59这个组,因为它的频率是3。虽然这个组似乎能够很好地代表这个分布的集中趋势,但是如果一个数值移动了位置,那么,众数则会发生很大变化。例如,如果第14个学生的分数是71而不是58。那么,分组的众数将是70—79,因为这个组现在的频率为3。你认为这个组能够代表这个分布的集中趋势吗?

众数的问题是它仅仅使用数据的唯一特征——出现最频繁的分数——来描述典型行为。它忽略了所有其他的分数。当你使用众数时,你扔掉了很多信息。例如,每个数值的顺序和大小。当你的样本量比较小,使用众数来描述数据存在很多风险。



PIE A LA MODE

中数(median),顾名思义就是处于中间的分数,比该数值大的值和比该数值小的值数量相等。在计算中数时,按大小顺序列举所有的数值,并且取处于中间的那个数。例如,在列举经济学专业的10个数值后,我们发现,第5个是62,第6个是63,所以中数是62.5。心理学专业的中数是56.5。中数不反应各个数值之间的差异,因为它只使用次序来计算。因此,我们可以改变一个分布的任意一个数值而不改变它的中数,只要处于中间的数值保持不变。同样,如果采用中数来描述数据,我们也会丢掉一些信息。

均数(mean),是所有数值的平均。也就是说,每个数值相加后除以相加数值的个数。例如,为了获得经济学专业的平均数,我们把10个数值相加之和为647,然后,除以10,得到的均数结果是64.7。心理学专业的均数是55.6。均数是一个分布的重心。因为平均数会受到每个数值的影响,改变分布中的任何数值均会改变均数。

哪一种方法能够更好地描述一个分布的集中趋势?与大多数有趣的问题一样,答案是“它取决于很多因素”。第一,这取决于分布的形状。如果是正态分布或者其他的单维对称分布,这三种方法得到的数值相同。但是,一个分布越偏态,这三种方法得到的数值就越分散。图12-3显示,平均数更容易受到分布右侧的极端数值大小的影响。中数受到的影响主要是因为分布右侧有更多的数值,而众数则不受到这些极端数值的影响[①]。你需要自己判断使用哪种方法。如果你想描述一个较大样本量人群的收入,你很可能得到类似图12-3中的分布。在这种情况下,中数可能是较好的选择,因为相对于均数,它较少受到少数具有很高工资的人的影响。你可以想象,一些非常极端的例子中,过大或过小的数值可能扭曲均数。因此,当你需要进行选择时,你最好检验分布的形状,通过分析这个集中趋势的使用目的,作出自己

---

① 选择哪一种指标也受到你所使用的数字特性的影响。见附录A中关于数字量表的讨论以及图A-1中哪一种集中量数更合适。

的判断①。

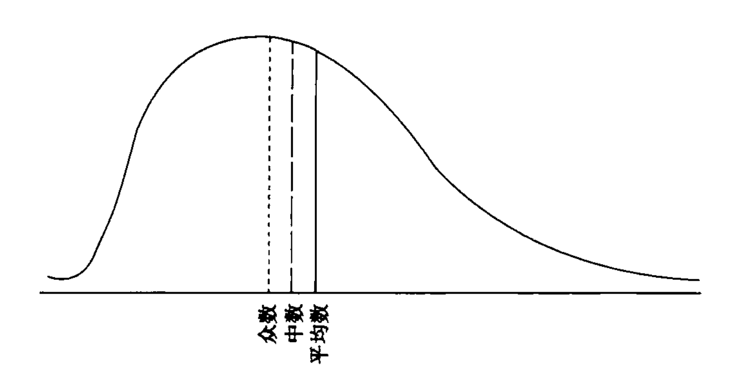


图 12-3 偏态分布中的众数、中数和平均数的位置

离中趋势

集中趋势能够告诉你关于分布的有用信息,但它只描述了数据的某一个方面的特性。第二个帮助你描述分布的统计方法是离中趋势,或者说这些分数的分散程度。

对离中趋势的一种测量是全距(range),用最大的数值减去最小的数值。在我们的例子中,对经济学专业来讲,全距是 91 - 46 = 45;对心理学专业来讲,全距是 74 - 36 = 38。虽然全距能够对离中趋势有所描述,但对多数数值是不敏感的,因为它只由最大数值和最小数值决定。一个极端的数值则可以完全改变全距。由于这个原因,另一种测量离中趋势的方法可能更有效。我们可以将每个数值减去均数,这样我们得到每个数值相对于均数的离散数值。为了得到平均离散数值,我们可以把这些离散数值相加并除以离散数值的数目。但是,因为这些数值会互相抵消,导致和为零,这就失去了意义。我们可以忽略离散数值的正负,将它们的绝对值相加,这样可以得到平均离散数值。但是统计学家发现,表示离中趋势的更好办法是把每个离散数值(也去除正负号)进行平方,将这些平方相加②,然后,除以相加的离散数值平方的数目。这样我们可以得到对离中趋势的测量,这称为变异(variance)。另一个更为有效的测量是变异的平方根③,叫做标准差。在附录 A 中可以找到计算这些数值的公式。

标准差是测量离中趋势的有效方法,这是因为它能告诉我们在标准差和均数之间存在多大比例的数据。对正态分布来讲,大约三分之二的数据落在均数以下一个标准差和均数以上一个标准差内。例如,如果我们让某大学一年级和四年级的学生进行写作测验,结果是一年级学生的标准差是 15,而四年级学生的标准差是 5,这样我们就知道,三分之二的一年级学生的成绩在 30 分之内变化,而三分之二的四年级学生的成绩则在 10 分以内变化。这样的结果以及四年级学生的平均分数相对高于一年级学生分数,因此,这可以为学校成功地教会学生写作能力提供证据。不仅个

---

① 我假设你已经阅读了第 5 章并试图公正地对待科学。Huff(1954)的《如何使用统计方法》和 Campbell(1974)的《统计的瑕疵与谬误》等书中给出了许多有关如何将像平均数等描述统计量曲解的生动的例子。

② 乘以它自身的数目。

③ 乘以它自身的数目得到的数字就是方差。

别学生成为了更好的写作者，而且学生整体的写作能力也得到提高。

你可能会发现，在使用均数作为描述分布的集中趋势时，用标准差来描述误差的大小是个有用的方法。事实上，均数是你能够描述任何个别分数的更好方法；一般情况下，标准差能够说明你的上述描述有多准确。如果所有的数值都是相同的，那么，标准差则为零，这说明使用均数不存在任何误差。随着数值间的距离逐渐增大，标准差也随之增加，你使用均数来代表该数据存在的误差也随之增加①。

描述变量间的关系

你完成了一个实验来检验自变量和因变量之间是否存在关系。虽然频率分布图是分析数据的首要方法，但是你可能会发现，描述实验变量间的关系图是个有用的方法。你对绘图应该并不陌生。在前些章中我们曾有一些讨论。但是，出于完整性的考虑，我们还是从基本的概念开始。

画 图

一个图通常有两个轴。竖直的轴（y 轴）称为纵轴，水平的轴（x 轴）称为横轴。在绘制实验结果时，将因变量绘制在纵轴上，自变量绘制到横轴上②。在某些情况下，自变量的水平不能用数值来表示，或者如果使用数值则没有量化的意义。在附录 A 中你可以看到这类水平被称为称名变量，因为它们就是简单的名字。在这种情况下，使用条形图来描述数据是比较合适的。图 12-4 是心理学专业和经济学专业的平均焦虑分数条形图。

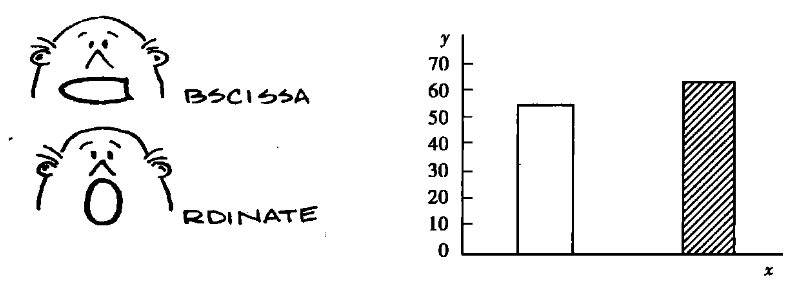


图 12-4 表 12-1 中心理学专业和经济学专业的平均焦虑分数的条形图

在很多情况下，自变量是连续的，或者使用附录 A 中的术语，这些变量的水平至少是可被排序的顺序变量。在这种情况下，如图 12-5 所示，你可以绘制直方图。直方图取消了条状图中各个条目间的空隙。图 12-5 是虚构的数据，是关于病人接受诊疗的时间与他们的自我图像评价间的关系。诊疗时间是个连续变量，因为我们可以在连续的时间上任意选择水平。

当自变量是连续变量时，描述数据的常用方法是使用线图或函数。图 12-6 展示了与图 12-5 相同的数据，但是使用的是线状图而不是直方图。首先画出单个的

---

① 选择哪一种离散指标也受到你所使用的数字特性的影响。见附录 A 中关于数字量表的讨论以及图 A-1 中哪一种集中量数更合适。

② 你也许已经发现，读坐标轴名称时的口型可以帮助你记住它们；"ab——"口型是水平的，而"or——"则是垂直的。这是我记住它们的窍门。

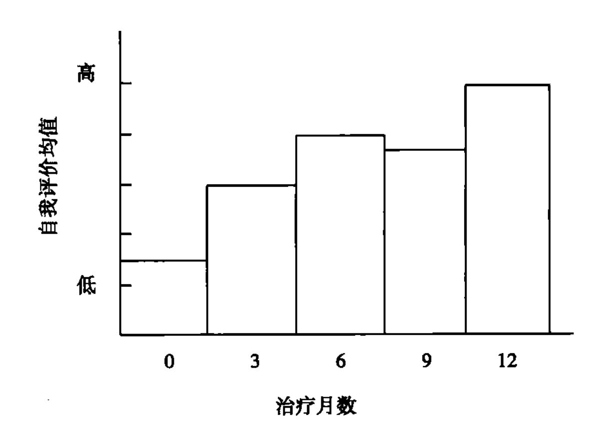


图 12-5 不同治疗时间的自我意向知觉的多水平实验结果直方图(假设的数据)

数据点,然后,用直线将它们相连。注意,这种描述数据的方法能比较有效地强调趋势。要使用这种图,你的数据必须是连续的。举一个差的例子来讲,假设图的横轴是种族的类别,如西班牙人、非洲裔美国人等,而不是接受治疗的时间。这些类别显然不是连续的,所以,列举这些类别的顺序也必定是武断的,试图发现这类数据的趋势不具有任何意义。

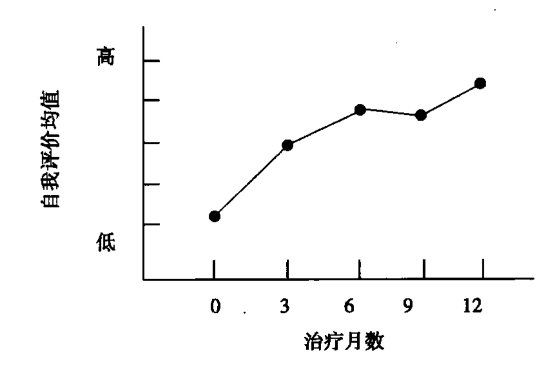


图 12-6 用线图表示 12-5 中的数据

线图在描述函数的(多水平的)实验而不是两水平的实验时显得更有效。两水平实验的问题在于,其实你不知道这两个水平间的关系是否是线性的,但你通常使用直线来连接两个数据点。对多水平实验而言,使用多于两个的自变量水平,这样即使你使用直线来连接相邻的数据点,你还是可以获得该函数的可能形状。但是,在解释函数的形状时要格外小心。函数的形状只有当图中的自变量水平是等距时才有意义。在附录 A 中,这样的水平被称为是等距的,因为变量间距是相等的。如果变量的水平是有顺序的(有顺序但是间距不等),那么,基本的变化趋势是可以被预测的,但函数的形状是不可预测的。(更多关于实验报告绘图的信息,见第 13 章。)

描绘函数

图 12-7 显示了几种类型的函数图。如果自变量每改变一个单位,就导致因变量在特定方向上改变等量的值,这种函数是线性的;任何其他函数都是非线性的。如果自变量的值增加导致因变量的值增加,这种关系称为正向的;反之,称为负向的。一个始终不改变方向的函数(也就是说,函数的方向始终为正或始终为负)是单调函数;否则,则是非单调函数。如果随着自变量的值增大,因变量的改变量逐渐增大,这种函数称为正向加速的;如果因变量的改变量逐渐减小,则称为负向加速。负向加速函数最终接近于一个特定水平并变得平坦。在这个位置上,曲线逐渐地越

来越接近一条直线，这条直线称为渐进线，不过该曲线和该渐进线是不会相交的。这种函数称为渐进函数或接近渐进线函数。

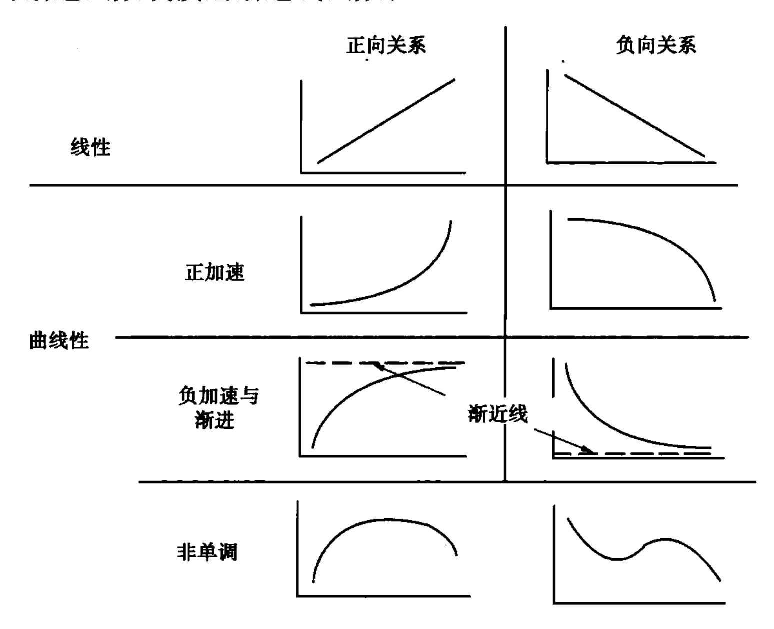


图 12-7 用于描述功能性关系的术语的图示

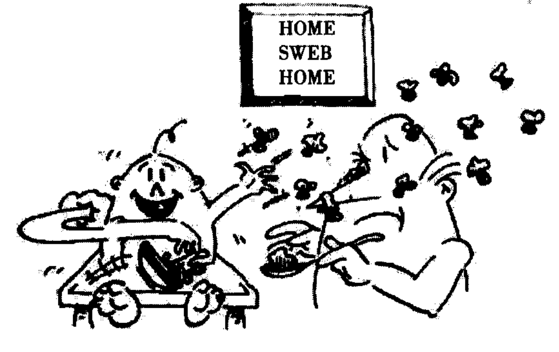
如果你是第一次看到这些术语，你可能觉得有点不知所措。但是，随着你使用这些术语来描述心理学关系的次数增加，你就会对它们越来越熟悉，并且它们能更有效地帮助你讨论结果。

描述关系的强弱程度

上面一节提到的或者是理想化的函数，或者是描述统计图，而不是单个数据点。但是，你很少在实际中发现每个数据都落在平滑的函数上。如果你使用原始数据来描述实验关系，你大概会发现这些数据围绕着某个函数周围变化分布。例如，在第1章中我们称之为散点图。

散点图

图 12-8 是几种散点图的例子。这些图可能是描述自变量和因变量关系的实验结果，也可能是画出了自变量和因变量的相关分析（第1章）。如果你观察到散点图中数据点的散落程度，你就可以设想这种关系的强弱。但是，视觉观察不是估计强度的有效方法。幸



一个密切关系的散点图？

运的是，当这种关系是线性的①，用一种叫做“相关系数”的描述统计值可以描述这种关系。

相关系数

相关系数是一个在 +1.0 和 -1.0 之间的数值，符号表示该关系是正向的还是负向的，大小表示该关系的强弱。相关为 1.0（+或-）表示一个很强的相关，而 0 则表示没有相关。

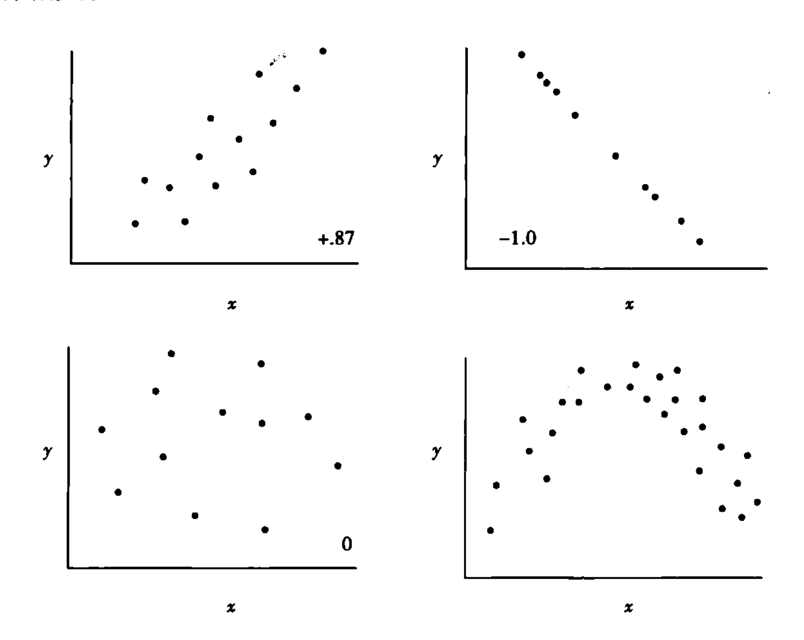


图 12-8 4 个散点图。3 个图中都标有相关系数，而右下角图中没有相关系数，因为其中的关系是曲线关系，可以用相关比率表示

图 12-8 显示了三组数据的相关值。右下角的图没有相关值，因为该函数显然是非线性的，使用简单的相关值是不合适的。[但是，有一种描述非线性关系的方法，叫做相关比（Kirk, 1990）]通过阅读附录 A 或阅读任何一本统计书，你会发现如何计算相关值②。

通常情况下，当你报告相关系数时，也会报告决定系数（变量被解释的比例）。简单来讲，决定系数就是相关系数的平方，它表示两个变量公共变异的比例。如果报告的相关值是 +.5，那么，决定系数是 +.25；两个变量共享 25% 的变异。很多研究者或者手册，包括美国心理学会发表手册（American Psychological Association, 2001）推荐，报告相关系数时，也要报告决定系数。但是，另一些研究者认为，相关系数是表述相关的更好选择，而决定系数严重低估了结果的实用意义（Rosnow & Rosenthal, 1999）。如果你使用这些统计方法，你应该熟悉这些争论，并采用合适的方法使用这些统计值。

---

① 事实上，在一种使用可以划分等级或顺序的数据的相关中，线性并没有什么意义。这种相关可以用于任何单调变化的关系中。

② 有些统计相关值都列在本章后。

解释析因实验的结果

析因实验(factorial experiment)的结果比其他类型的实验结果更难解释,因为它使用的自变量数量多于1,并且需要对交互作用进行解释。图12-9的上部显示了前面我们的一个虚构的实验数据,其中我们测量了学生阅读字体为12号或10号的文本时的速度。在这个例子中,假设我们使用了8岁学生为一组,12岁学生为另一组。注意,一个自变量(字体大小)被画在了横轴,另一个自变量(年龄)采用点和线型来表示。我们不再简单地问自变量的变化是否对因变量产生影响。我们需要问三个典型的问题:①字体大小是否有效应?②年龄是否有效应?③一个自变量的作用是否取决于另一个自变量的不同水平?前两个问题指的是主效应,第三个问题指的是交互作用。

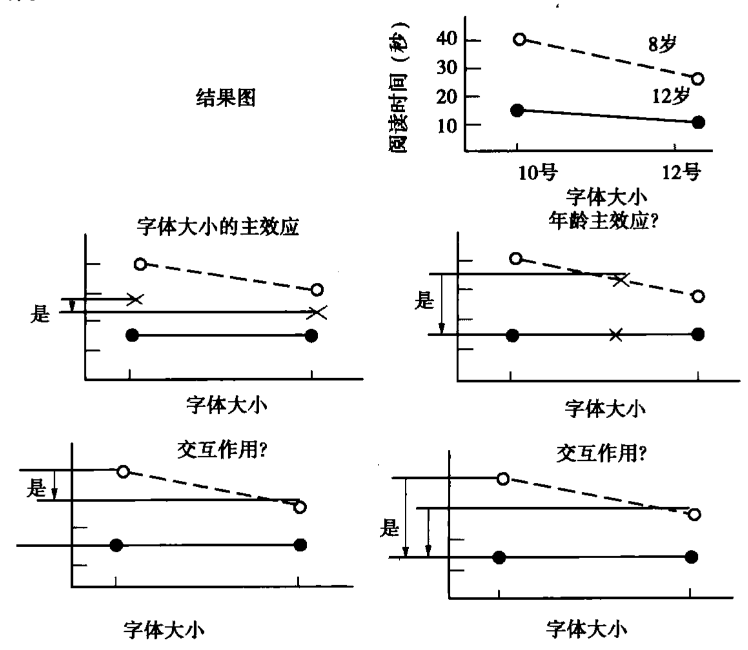


图12-9 一个2×2的析因设计中的主效应和交互作用

主效应

因为我们在该实验中使用了两个自变量,因此,有两个可能的主效应。为了揭示主效应是否显著,我们需要进行统计分析。为了解释字体大小实验的结果,我们假设任何我们可以看到的效应都是显著的。为了确定是否存在字体大小的主效应,我们需要忽略年龄的任何效应。这样,我们需要在图上对每一个字体大小的线上找到一个点,这个点处于两个年龄组数据点的中央。在图12-9中"字体大小的主效应?"下面,你可以看到×处于对应的空心圈和实心圈的中央位置。这些×代表在平均了年龄的情况下,字体大小产生的效应,就像是年龄根本没有被操纵。为了确认两个×是否存在阅读速度差异,我已经画出了与纵轴相交的水平线。箭头显示,代表字体大小的这两个×之间存在差异。所以,对上述问题的回答是肯定的,即存

在字体大小的主效应。现在使用同样的步骤，你可以确定是否存在年龄的主效应。

交互作用

如果你不记得什么是交互作用，你可以回想一下第 9 章中关于析因实验中交互作用的讨论。请记住，当一个自变量的作用取决于另一个自变量的水平时，便产生了交互作用。在上述的字体大小实验中相应的问题是，字体大小的效应是否受到学生年龄的影响？或者说，年龄的效应是否受到字体大小的影响？

在图 12-9 的左下部，我画出了与纵轴相交的水平线来检验当字体从 10 号变到 12 号时对每个年龄组产生的效应。请注意，对 8 岁组而言，阅读时间减少；而对 12 岁组来讲，时间没有变化。因此，字体的变化产生的效应确实取决于年龄段，所以，对上述问题的回答是肯定的，即存在交互作用。在该图的右下部，我采用另一种方式提出问题，“年龄对阅读时间的效应取决于字体大小吗？”如果你根据图中的线进行判断，你会再次看到存在的差异。当然，只存在一个交互作用，所以，对上述两种问题方式的回答总是一样的。但是，采用两种方式检验交互作用能够帮助你加深对它的理解。

图 12-10 是这个实验的其他可能结果。使用我们刚刚讨论的同样步骤，我们可以对每个图回答这三个问题。

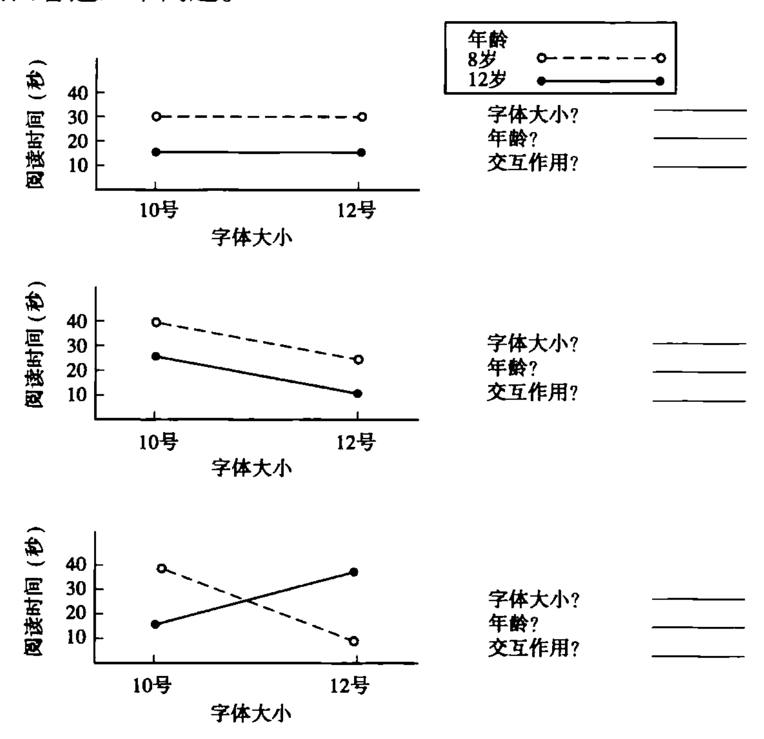


图 12-10 一个 2×2 的析因设计的可能结果。对每个图回答三个问题（答案可以在本章最后找到）

在试图检验这些图中是否存在主效应和交互作用时，我希望你注意到，当只存在主效应时，是否存在交互作用是比较明显的。但是，当存在交互作用时，对主效应的解释比较复杂。例如，对于图 12-9 中的结果而言，虽然存在交互作用，但年龄对阅读时间的主效应是有意义的，因为年龄的效应在两个字体大小均存在。但是，虽

然字体大小也存在主效应，由于交互作用的存在使得该主效应存在问题，因为该效应只在一个年龄组存在。图 12-10 的底部展示了称为相交的交互作用（crossover interaction）的情况，即两条线是相交的①。当存在相交的交互作用时，主效应比较难以解释。事实上，此时主效应没什么意义了。

目前的讨论局限于最简单类型的析因实验，其中只有两个因素，每个因素有两个水平。如果再增加一个水平或增加一个因素，解释主效应和交互作用将更加困难。为了使你体验一下这种难度，假设我们在阅读实验中增加一个因素，即阅读材料的难度：难或易。图 12-11 是可能的结果。在这种情况下，你需要问，这三个因素是否存在主效应：年龄、字体大小和难度。另外，你还需要问，年龄是否与字体大小交互、年龄是否与难度交互以及字体大小是否与难度交互。上述任何两个因素间交互称为二维交互作用，因为它涉及两个因素。为回答这些问题，你需要平均第三个因素引起的效应并找到平均值，就像我们之前试图找到主效应那样。然后，你就可以像前一节讨论的那样来解释二维交互作用。

在这个例子中，还存在三维交互作用。当任何一个二维交互作用取决于第三个因素的水平时，便出现了三维交互作用。在这个图中，年龄和字体大小的二维交互作用是否受到阅读材料难易的影响？如果是，那么，便存在三维交互作用。你可以看到，当自变量多于两个时，解释交互作用将变得困难，虽然解释实验结果的基本步骤是不变的。

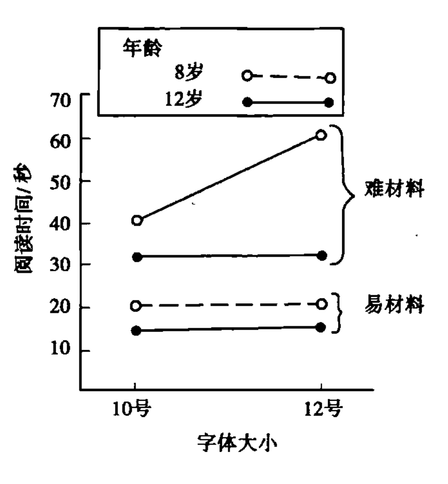


图 12-11 一个 2×2×2 的析因设计中的三维交互作用

推论统计

为了讨论推论统计的一般逻辑，让我们回到心理学专业和经济学专业学生焦虑分数的例子上。为了检验这两组的焦虑分数是否存在差异，我们对每个组绘制了频率分布图并计算出平均值。我们发现，经济学专业的均数是 64.7，而心理学专业学生的均数是 55.6。那么，差异真的存在吗？你当然可以说：这个差异不是差异吗？对这两个样本来讲，你当然是对的；两个样本间的任何差异都是真正的差异。但是，心理学家问这个问题的本意不是“你在这个实验中随机选择的两个组的分数是否存在差异”，而是“在你选择抽样的整个心理学专业和经济学专业群体的焦虑分数是否存在差异”。实验的目的是对这两个抽样的总体进行检验，而不是对特定的样本进行检验。

假设你是一个豆农。由于豆子虫害导致庄稼收成不太好。豆子虫害是一种严重的灾害，能够导致很多豆子枯萎凋零。为了检验你能否去除豆子虫害，你在一块新土地上种植了一种新型的抗病虫害豆子。在收获了被豆虫害危害的田地和新田

① Rosenthal and Rosnow（1981）称这种相交的交互作用是仅存的交互作用，因为他们认为，在画图和揭示之前应将主效应剔除，剩下的只是相交的交互作用。

后，你得到了两箱豆子，每箱有 10 吨。你准备确定这些豆子是否有病害。[①] 当然，你不想对所有豆子进行检验。所以，你决定从每个箱子里拿出 100 个豆子作为样本。你发现，在被豆子虫害危害的样本中有 12 个枯萎的豆子，而另一个样本有 7 个。显然，样本间存在差异，但你想知道的是这两个不同的豆子总体是否存在差异。推论统计可以帮助你回答这个问题。"推论"指的是这种检验能够帮助你推测这两个总体间是否存在差异。

你作为一个心理学家，与上述的豆农面临的问题很相似。你已经从两个可能存在差异的总体中（也就是自变量的两个水平）选择了样本数据，你想知道这两个总体的行为是否存在差异。

虚无假设显著性检验

很多推论统计是基于虚无假设的显著性检验。我在前些章中简短地介绍过这个概念，下面详细进行介绍。虚无假设是指自变量的水平变化不产生效应。例如，假设你想知道男性和女性在数学能力上是否存在差异。你可能会让男性样本和女性样本解答 SAT 考试中的数学部分来回答这个问题。虽然你做这个研究的原因是你认为存在这样的差异，显著性检验的步骤要求你相信这两个组不存在差异。假设你的结果表明，男性样本的均数和女性样本的均数存在差异。当你进行推论统计检验时，你所问的问题是，如果虚无假设是真的，也就是说男性和女性整体上不存在差异，那么，在多大的可能上，样本均数存在的差异是由于抽样误差造成的？

因为样本通常是随机地从总体中进行抽取，那么，即使两个总体不存在差异，你也可能发现样本间的大小差异。这样的结果在样本量小的时候通常更明显。推论统计将样本大小等因素也考虑在内，并告诉你，当虚无假设为真时，在多大程度上你会得到你的结果。

显著性检验中的决策误差

无论任何时候我们都应该在证据的基础上作出是否判断时，我们均有一半的几率或对或错。假设，从测谎仪上你得到了有关数据，你需要对被检测者是否说谎作出判断。以下情况你是对的：你认为她说的是真话，而她确实说了真话；或者你认为她说谎，而她真的说了谎。但是，你也可能犯错：你认为她说谎，而她说了真话；或者你认为她说了真话，而她其实说谎了。表 12-2 列举了对虚无假设检验的结果。

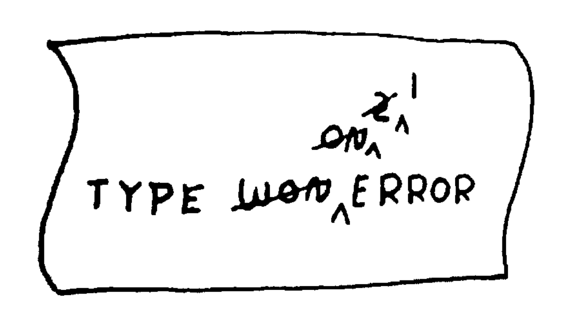
表 12-2 在虚无假设检验过程中可能的正确或错误种类

	真实	
	虚无假设为假	虚无假设为真
拒绝虚无假设	正确（功效）	Ⅰ类错误
无法拒绝虚无假设	Ⅱ类错误	正确

如在表中列举的，从统计检验的几率结果中，你可以得出结论，或者拒绝虚无假

① 快速说三次。

设或者不去拒绝它。(注意本章稍后将讨论,你不能说应该接受虚无假设。)当你得到上述任一结论时,你可能是对的,也可能是错的。你错的方式有两种名称。当你拒绝了一个真的虚无假设时,你犯了Ⅰ类错误;当你没有拒绝一个假的虚无假设时,你犯了Ⅱ类错误。统计检验告诉你,你在多大程度上犯了类型Ⅰ错误,它被称为显著性水平。



显著性水平

虽然我们无法摆脱这样严格的标准实属不幸,但大多数心理学家一致认为,如果一个结果是显著的,那么,从样本中观察到的差异是由抽样误差引起的概率应该小于1/20。这样,如果样本确实是从同一个总体中抽取出来的,那么,你只可能在20次中有1次(或者100次中有5次)得到显著的差异。一些心理学家更加小心,避免当总体不存在差异时却声称发现了差异。只有当检验显示,100次中有1次是因为抽样误差造成了差异时,他们才认为,这反应了真正的差异。这些策略被分别称为在.05水平或者.01水平上进行的显著性检验。如果达到或超过了这些检验概率,那么,结果就被称为在统计上是显著的。

当你阅读文献时,你会发现这些显著性水平用 $p < .05$ 或 $p < .01$ 来表示。这个术语表示,该检验在.05或.01水平上是显著的。你可以认为,自变量水平上的差异在100次中有5次,或者100次中有1次可能是它们实际上来自同一个总体。请确认符号指示的方向;$p > .05$ 表示该检验不显著。

还需要注意的是,我对自己陈述的措辞。一些学生错误地认为,当你在某个显著水平进行检验时,你可以说在多大程度上你的结果可以重复。例如,如果你在 $p < .05$ 水平上发现了一个显著统计结果,你可以说,如果你重复这个实验,那么,拒绝虚无假设的概率是95%。但是,你不能从统计显著性的水平得出结论说,你在多大概率上能够重复你的结果!

统计功效

我们已经讨论了在多大概率上你可能得到错误的推论,但现在假设我们是正确的。有什么办法能知道多大可能我们是对的吗?从表12-2中你可能已经注意到,我将"功效"这个词放在了临近正确拒绝错误的虚无假设的括号内。效力指的是统计检验允许你正确拒绝虚无假设的概率。虽然这超出了本书的范围,但你应该知道,有可能通过计算统计检验的效力来确定多大程度上你会错过自变量引起的真正差异,也就是说虚无假设为伪(差异真正存在)。例如,如果检验的功效被定在0.50,我们知道,你做这个实验无法拒绝虚无假设的概率有50%。影响功效的三个

因素是显著水平、潜在效应的大小以及样本大小。其中,研究者能够更多控制的是样本大小。所以,研究者经常在做实验之前计算功效,从而检验他们使用的样本是否足够大。

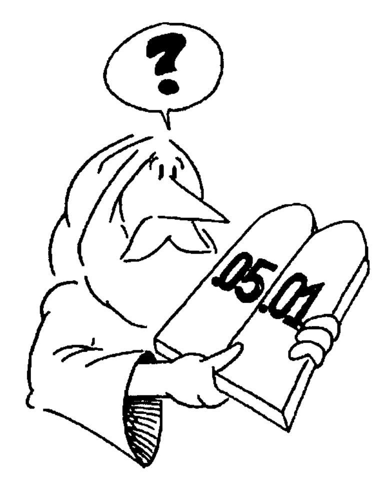
参数检验和非参数检验

有很多推论统计可以帮助你作决策。你的选择取决于实验设计以及数据符合何种检验假设。(见附录 A 中一些推论检验的样例。)最常用的检验叫做参数检验。在这种检验中,假设是样本总体的频率分布是正态分布。当这个假设不能被满足时,你必须使用非参数检验。

推论检验很显然是评价心理学实验结果的重要工具。事实上,复杂的统计检验的发展是推动心理学成为科学的重要因素。但是,我们必须认识到推论检验的局限。

对统计检验的误解

有研究者认为,当统计检验没有显示出自变量水平的显著性时,这说明,这些水平是显著相同的。这是错误的! 为了避免这个错误,我们应该记住的是,推论检验是用来告诉我们,如果样本来源于同一个总体,那么,你得到差异的概率;而不是告诉我们,如果样本来自不同总体,那么,你得到没有差异的概率。从而,阴性结果(统计不显著的结果)很少在心理学杂志上发表。总之,统计检验不是为了告诉我们,当两个样本来自不同总体时,它们表现出等价的概率;反之,检验只告诉我们,样本来自相同总体的概率①。



研究者易犯的第二个错误是,当使用推论检验时,他们将.05 和.01 视为教条;他们如果肯注意.06则不会走死胡同。对待显著水平更现实的做法是,还原它们的本来面目,即帮助你作决策的工具。无论何时你需要作出具有不确定性的决策,你不仅应该考虑正确或错误的概率,还要考虑正确或错误的收益和损失。在其他的决策中,你不能忽略这些因素。例如,如果你要决定是否乘坐飞机,相对于要决定是否带伞时,前者让你需要高概率的晴朗天气。收益和损失在这两种情境下很不相同。.05 还是.01 水平不影响这些收益和损失。因此,在解释你的实验结果时,你应该考虑正确或错误的后果,而不是盲目地锁定在.05 水平。

在心理学界存在一些使用"显著性"变式(例如,"非常显著")的争论。有人认为,使用类似的变式是错误的,因为心理学的传统是区分显著的或者不显著的结果。

---

① 尽管我们的统计检验不是检验相同的观点在大多数心理学文献中得到印证,但它并不是放之四海而皆准的真理。正规的统计学对相等性的检验已经发展了数十年,虽然心理学家对此并不熟悉。这些内容超出了本书范围,但我依然给出有关参考书目"Using significance tests to evaluate equivalence between two experimental groups. Rogers, Howard, and Vessey(1933)"以供感兴趣读者参考。

更重要的是，使用这种变式会错误地将效应大小与效应概率的大小相混淆（Harcum, 1989; Levenson, 1990）。但是，另一些人则认为，因为概率是连续的，说一个效应更显著并没有什么错（Kanekar, 1990）。争论的双方可能都同意，他们避免将统计显著的水平混淆为实际的显著性——我们刚刚讨论过这个问题。所以，当报告一个阳性结果时，最好说明这是统计显著的，以此说明你并没有必然地认为，该结果是实际显著的。当报告阴性结果时，说明该结果是统计上未显著（statistically nonsignificant）而不是不显著（insignificant）。不显著这个词显然暗示该结果不重要。

当你尝试避免混淆统计显著和实际显著时，请记住这句古语——当差异存在意义时，差异才称之为差异（a difference is a difference only if it makes a difference）。假设你是个雇主，“速指快速阅读（Fast-Finger Speed-Reading）”公司试图说服你雇佣他们来训练你的雇员进行快速阅读。他们说，他们有实验证据证明，参加他们课程的人阅读速度明显加快。出于好奇，你问他们速度提高了多少。他们承认，研究显示他们的学生平均每分钟加快一个字，但他们坚持认为，该结果是统计显著的。他们也许是正确的。通过招募足够的被试和收集足够的数据，即使是群体的微小差异也可能是显著的。但作为一个雇主，你更注重实际意义而不是统计意义。作为一个科学家，你也应该如此。

为了鼓励研究者更多地考虑他们的结果是否有意义并且在统计上显著，而不是教条地在.05 水平上进行检验，一个重要的心理学杂志的前主编 Geoffrey Loftus（1993, 1996）指出，统计假设检验通常是没必要的。他鼓励研究者呈现的数据应该包含均数和相关的离中趋势值（如标准差）。他相信，在很多情况下，检查这样的图就能够马上明确地告诉你效应的显著性，而不需要使用推论统计检验。如果是这样，他不鼓励使用这样的检验。

最后，评价差异的实际意义是判断结果重要性的关键。当你的结果是重要的，那么，本章中讨论的工具能帮助你确定其重要性，但这些工具不能建立结果的重要性。作为研究者，你必须通过逻辑推论说服其他研究者，你发现的差异是有意义的。

元分析

虽然你可能不会在简单实验中使用元分析，但你将会看到在很多文章中提及这种统计技术。你应该了解元分析以理解这些文章。在第 6 章中，我们讨论了如何搜索文献。当你进行搜索时，你会惊奇地发现，围绕某个特定主题的文章如此之多。即使是相对某一狭窄的领域，其文献库也会包含成百上千的文章。研究者们在撰写综述以评价和整理这些文献时所采用的典型方法是叙述性的。研究者略读了所有文章，分类出一些重要的研究，其他的则被认为不那么重要。他将尽量记住所有重要的研究结果，然后，总结出主要的发现。研究表明，研究者们完成这一流程的方式是很草率的（Jackson, 1980）。在很多情况下，不同的研究者会从相同文献中得出完全不同的结论。问题在于这些研究者们面对的是几乎不可能完成的任务，就如同强迫他们看某个实验 100 个被试的数据，然后，在不使用统计分析的情况下给出结论。元分析提供了一种统计学手段，用来整合大量不同研究的数据。

你无需掌握元分析的统计细节就可以理解它的分析结果。如果你确实需要知

道更多细节，可以参考在本章末尾列出的有关元分析的书籍。简单来说，元分析使你能够对研究同一问题的实验结果进行整合，即使这些实验使用不同的方法也没有关系；然后，用统计学方法整合他们。例如，Lipsey 和 Wilson（1993）曾研究过心理的、教育的和行为的治疗是否有效。他们查看了 15 年间有关此课题的 302 项研究。在元分析的审查中，每个研究的基本数据是其平均处理效应大小（mean treatment effect size）。该统计值很容易计算，它是治疗组的平均值减去控制组的平均值，然后，除以控制组的标准偏差。所以，在 Lipsey 和 Wilson 的综述中，无论各研究的因变量如何测量或者如何实施治疗，我们只是用治疗组的平均值减去不施加治疗组的平均值，然后，这个数再除以后一组的标准偏差。通过对这些平均处理效应作为数据的分析，作者判断这些效果有多大可能性出自偶然。这些数据也可以用各种方式重新分析。例如，所有采用了某个特定实验设计的研究都可以单独分析；或者将高质量研究和低质量研究分开分析。这种小范围的分析使得研究者们可以估计，对所有研究的初始分组是否合适。

有些研究者对元分析持批评态度。Wilson（1985）认为，该技术有时被用于整合很多有严重瑕疵的实验结果。因此，元分析的结果就会和那些原始实验一样带有瑕疵。然而，元分析也不乏支持者（Mann, 1990），支持者认为，包含了各种不同分析的成熟研究可以使从有瑕疵的实验合并出一个有超级大瑕疵结果的可能性变得最小。无论如何，元分析肯定能够生根、发芽。如果运用得当，它将是整合大量臃肿的研究结果的宝贵工具。

借助计算机解释研究结果

计算机被应用于心理学实验的很多领域，如文献搜索、刺激显示和记录反应等。然而，计算机应用最广泛地是数据的统计分析。计算机之所以对完成此类工作特别有价值，是因为它们能够快速地存储和处理大量数据。近年来，计算机的可用性进一步增强，如今强大的统计软件使得大多数的统计测试都能在个人电脑或者笔记本系统上完成。

尽管计算机在统计计算方面的效果一直都是绝对正面的，但在计算机的使用中还是存在一些问题。一个问题是，在统计工作中，计算机减少了一些手工操作，使得研究者好像不需要付出很多努力。但在一开始使用它们时，理解统计检验过程至关重要。计算机却允许你跳过这种理解过程，只要有人给你演示输入数据的必要步骤，然后，计算机就会给出结果。那么，你就有可能在不理解做了什么的情况下完成所有流程。我要求学生在使用计算机前手工计算每个统计检验，这样他们就会明白在那神秘的盒子里到底发生了什么。

同时，计算机非常完美，它几乎不犯错误，这使我们轻信它给出的结果都是正确的。但常言道，输入错误则输出自然也不正确。你应该记住在本章中学到的关于解释结果的教训。此外，你需要知道你可能使用的各种统计方法的假设条件和局限性（见附录 A）。最后，你不应该在进行一些简单的验证前就无条件接受计算机输出的结果。尽管计算机不大可能犯错，但你可能在设定和录入数据时出问题。计算机快速而准确，但它也极端愚蠢。计算机一点都不关心你是否犯错，即使你不小心把本

应按照先条件 A 后条件 B 顺序输入的数据倒了过来。计算机不会检查你是否犯了此类错误,你应该自己检查。

有一种方法可以用来快速检验你的统计结果,即查看计算机提供的描述性统计结果,看它们是否有问题。例如,在我们字体大小的实验中,我们希望 8 岁孩子的阅读速度比 12 岁的慢很多。如果这种符合逻辑的预期没有被满足,我们就应该对计算机的分析保持警惕,这可能意味着我们在向计算机输入数据时犯了错误。我们也可以手工计算一部分数据的平均值,以检验手工结果与计算机的结果是否一致。在字体大小实验中,我们就可以计算一个年龄组的、一段的,或者特定字体的平均值,用所得平均值与计算机输出结果中的对应数值进行比较,看它们是否一致。少量的简单验算只需要很少的时间,却可以使我们对结果的正确性更加自信。

我希望我已经把计算机的角色讲清楚了。计算机和统计工具软件只是可以使数据解释工作简单化的工具。毫无疑问,电脑会让你不再恐惧。它们是你的朋友,而且日益友好。但是,如同所有复杂工具一样,你必须小心以确保电脑的正确使用。这些电脑朋友是完全不知道变通的,你需要严格遵循它们的规则。它们相信你说的一切,哪怕是错误的;而且它们也没有常识来判断你何时犯错。要避免麻烦,你需要了解它们的能力,也需要了解它们的局限。

小　结

一旦你完成了一个实验,你就必须解释得到的反应数据。最开始,一种有用的方法是画出一个频率分布图,以描述变量不同水平上各个数据的数量。有时这些分布是钟形的对称分布,它被称为正态分布。双模态分布有两个频率最高的峰;偏态分布在某一方向的尾巴上有更多数值;截断分布则在某一方向上的尾巴是缺失的。有三种重要的统计值来描述集中趋势。众数是出现次数最多的数值,中数是处于中间的数值,而均数则处于分布的重心。有两种统计值用来描述离中趋势。全距是最大值和最小值之间的差异;标准差(有时称为变异)描述了接近正态的分布的离散程度。

图能够描述自变量和因变量之间的关系。自变量的水平绘制在水平轴上,即横轴;因变量的值绘制在竖直轴上,即纵轴。条状图用来描述在量上不同的类别数据。直方图或函数线图都可以用来描述连续变量。在描述函数时,你可以说明它们是线性的或非线性的,正向的或负向的,单调的或非单调的,正向加速的或负向加速的,或者渐进的。实验关系的强弱可以在散点图中表示,如果关系是线性的,你可以计算相关值或决定系数。

为了解释析因实验的结果,你必须确认主效应是否存在,主效应指的是当另一个因素的效果被平均后,一个因素对因变量的效应。另外,你还要确认一个变量的效应是否取决于另一个变量的不同水平,这样的差异称为交互作用。另外,相交交互作用会使得主效应的解释变得困难。当实验有三个或更多因素时,解释交互作用变得更具有挑战,因为可能存在多个二维交互作用以及三维交互作用,或者更多因素引起的更高级别的交互作用。

推论统计用来推断在多大程度上样本的差异是由于误差造成,而不是由于总体

（自变量的水平）的真正差异造成的。在统计检验中，设定一个虚无假设用来说明自变量的水平没有效应。如果虚无假设为真，统计检验可以确定数据样本中出现的差异是由误差造成的概率。一旦统计检验判定了虚无假设是否为真，就可能发生两类错误。Ⅰ类错误指的是当虚无假设为真时却被拒绝了；Ⅱ类错误指的是当虚无假设为假时却没有被拒绝。当虚无假设为假时，它应该被拒绝，统计检验能够正确执行此拒绝过程的概率称为统计功效，这取决于实验的敏感性和被试的个数。如果一个效应在统计上显著，那么，该效应是由于抽样误差造成的（也就是说Ⅰ类错误），概率应该小于.05 或 .01。参数检验假设总体分布是正态的，而非参数检验不需要这样假设。研究者有时候错误地使用了统计检验，如他们认为，不显著的结果标志着条件的同一性，他们过分地强调.05 和 .01 的显著水平，或者混淆了统计显著性和实际显著性。

元分析是一种整合多个实验结果的统计方法。对每个实验计算单个的统计值，这被称为平均处理效应大小；然后，分析这些效应大小，以确认这样的效应在多大程度上是由于误差造成的。

我们常常可以用计算机来进行统计分析。但是，你必须小心地保证它满足统计检验的假设，并且数据的输入方式正确。在接受该结果前，应该检查结果的内部一致性和正确性。

图 12-10 中的问题答案：

上图：	字体大小？	无
	年龄？	有
	交互作用？	无
中图：	字体大小？	有
	年龄？	有
	交互作用？	无
下图：	字体大小？	无
	年龄？	无
	交互作用？	有

推荐的统计书

初学学生

Hinkle D E, Wiersma W & Jurs S G. (1988) *Applied statistics for the behavioral science*. Boston: Houghton Mifflin.

Kirk R E. (1990). *Statistics: An introduction*. Fort Worth, TX: Holt, Rinehart & Winston.

高年级学生

Keppel G & Zeddeck S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches*. New York: W H Freeman.

Maxwell S E & Delaney H D . (1990). *Designing experiments and analyzing data: A modle comparison perspective*. Belmon CA: Wodsworth.

Myers R H. (1971). *Response surface methodology*. Boston: Allyn & Bacon.

元分析的推荐书目

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, et al. (Eds). (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Glass G V, McGaw B, & Smith M L. (1981). *Eta-analysis in social research.* New bury Park, CA: Sage.

Rosenthal R. (1991). *Meta-analytic procedures for social research* (Rev. ed). Newbury Park, CA: Sage.

13

如何报告实验结果

自我满足是写出好文章的大敌……，与组织优美的语句以及整理有说服力的论据一样，批判性地阅读自己的文章也是一项十分重要的技能。

RACHEI TOOR(2006)

在真理面前我们都是盲目的探索者，
被杂乱的只言片语所迷惑，
我们是否说出了本意，
或清楚表达了所说的内容，
我们几乎不知可否，
只有时间才能检验对与错。

B. DECKER(1967)

众所周知，《十诫》已被公众所接受且出版。然而，有一点你可能不知道，摩西要求继续修缮剩下的 34 条；但是，由于延迟出版，还有超过 16 条的内容仍未被世人所知。

PALLADINO & HANDELSMAN(1995)

已出版的内容都是经过严格审查并通过理性判断的研究，这些研究中的内容符合标准构架。尽管老练的研究者们可能通过字里行间读出真正的含义，但在文章中却并未真正包含这些内容。

MADIGAN. JOHNSON & LINTON(1995)

有这样一个经典的哲学争论,森林中有一棵树倒了下来,要是没有任何人听到这倒下的声音,那么,树倒下时是否发出了声音?问题就在于有人是否听到了能够被称之为声音的声音。你怎么看这个问题呢?我们会在研究报告中提出相同的问题,即没有人听说过的研究算是研究吗?形而上学的回答是,你怎样界定这一问题。我们非常关心现实情况,至少可以认为,未报告的研究可能是尚未研究过的。研究的最终目标并不在于做实验本身,而是在于建构科学的知识体系。如果其他的科学家不了解你的实验,那么,你的结果就不能成为组成科学知识的基本元素。通过实验报告将你的研究结果公布于众,科学事业才可以从你的研究中获益。

实验报告是你的研究成果,因此,你应该尽力呈现一份高质量的成果。一份文笔优美的实验报告无法挽救一项糟糕的研究;而一份文笔糟糕的报告也会毁掉一项好的研究。据我了解,如果通过非正式的讨论对某些研究进行评判,研究者们就关键性问题所作的充分的实验就会得到认可。但是,如果通过论文形式呈现,由于他们的书写能力相当贫乏,以至于他们的研究成果无法被外人知晓。很多优秀的研究都可能由于这个原因而不为人所知。

即使是写作课的老师也难以教授学生该如何有逻辑地表达想法。在本章中我没有足够的篇幅教授你怎样写作①。我所见过的最简洁的写作方法是 William Safire (1979)提到过的:

切记:绝不可以"分离"不定式。不要使用被动语态。不要以消极的方式阐述观点。动词与主语相一致。仔细校对,以防词语遗漏。通过重读文章,你会发现重新阅读和编辑可以避免很多重复之处。作者不可以中途改变观点。不要用连词作为一个句子的开始。不要过度使用感叹号!!! 代词与指代词的距离尽可能地近,特别是在长句子(十个单词或者更多一些)中。写作时要仔细,不可使用垂悬分词。在句子结尾处使用系动词是不恰当的。不要使用含混的隐喻。不要使用听起来显得古怪的流行词汇。在文章中,代词必须与所替代名词数量相一致。正确使用习语。副词放在动词之后。还有最后一点很重要:不要过度使用陈词滥调:寻找可行的代替办法。

这一章的目的有限。我会对研究报告的构成进行介绍,对判别文章的可读性提供一些建议,还会提供一些带注解的报告范文。

研究报告若能遵循一致的格式,将有助于信息的有效传递。为了便于记忆,美国心理学协会(American Psychological Association, APA)(2001)专门为了研究报告的写作编纂了一系列写作规则——《美国心理协会写作手册》(*Publication Manual of the American Psychological Association*)②。这本手册应该成为每位实验心理学研究者的必备工具书。本手册最初出版于 1929 年,当时只是一个 7 页的期刊论文,到现在已经成为了一本 439 页的书。正如你想象的那样,要掌握 APA 格式中的所有内容要花费一定的工夫。然而,下工夫阅读这本书是十分值得的。这本写作指南是一份被广泛使用的标准参考。比如,该书的出版商就要求每位作者使用这一指南。除非

---

① 如果你写作有困难,请看一看 Strunk 和 White 的 *Elements of Style* (1979),也许会有帮助。

② 写研究报告时,确保使用第五版的 *APA Publication Manual*。第五版相对第四版有所改变。从 287 到 299 页,我使用了一个报告范例列出了这些变化,但你还是需要进一步仔细阅读参考手册。

在其他课程上导师专门为你指定了一份格式手册，否则，你可以很放心地使用这一手册中的规范。在心理学课上，即使是一份班级作业，你也应该按照 *Publication Manual* 中的要求来书写。尽管我无法将这一册书中的所有内容都进行介绍，但是，我会列出最重要的规则，还会指出研究新手们经常会犯的错误①。

APA 格式与其他写作格式的不同

在我真正阅读 *Publication Manual* 的规则之前——这些规则统称为 APA 格式——我想要探讨一下 APA 格式与其他写作格式存在哪些不同。心理学中的一些规则可能不是你马上就能清晰掌握的，有时这些规则只在 *Publication Manual* 中才有所讲解。心理学专家们有时候也很难区分出这些写作格式有何差异，因为在阅读了大量期刊类文献之后，这些格式已在他们头脑中形成了根深蒂固的印象。对于学生而言，其他课程的写作格式可能与心理学不同，因此，意识到这些细微的写作差异十分重要。我所提到的很多内容都是以 Madigan，Johnson 和 Linton（1995）的研究为蓝本，他们比较了两份心理学杂志《现代语言协会出版物》（*Publications of the Modern Language Association*）和《美国历史杂志》（*Journal of American History*）中语言使用的情况。

语　言

心理学研究者们试图使语言清晰易懂，也就是说，用语不应该妨碍到信息的表达。在人文学科中，文章中的用语和思维通常以某种方式相联系，即词汇的选择与思维的呈现同等重要。在心理学文献中，用语要尽可能简单、浅显易懂。例如，虽然 *Publication Manual* 不赞成使用被动语态，但很多科研论文还是习惯采用被动语态，而很少用主动语态。于是，就会出现“数据被分析（the data were analyzed）”，而不是“我分析了数据（I analyzed the data）”这样的说法。尽管这看上去只是一些细微的差别，但被动语态并未把强调的重点放在“研究者”上，而是放在了“数据”一词上。如果你想要在科研文章中以一种新颖、独特的方式论述某事物，其他的研究者可能并不领情，他们还会因这种方式质疑你写作的严肃性。

引　用

心理学者与历史学者及文学评论家引用文献的方式不尽相同。人文类作者通常会直接引用。平均而言，历史学者们每 60 个单词中就会出现一次直接引用，然而，心理学者们每 3 000 个单词中才会出现不到一次直接引用（Madigan et al, 1995）。心理学家们不会直接引用，而是转述作者的意思。正如前面所提到的那样，这一差异可能反应了语言使用的差异，历史学家认为，语言的表述方式和表述的内容同样重要。但心理学家却认为，数据、理论和论据是独立的，不依赖于所使用的词

---

① 如果想真正掌握 APA 格式，你可以参考美国心理协会的一本书“*Mastering APA Style: Student's Workbook and Training Guide*”（2002）。或者另外一本“*Mastering APA Style: Instructor's Resource Guide*”（2002）。在 http://www.apastyle.org 中可以找到这些资源以及其他一些可用的资源（如“*Concise Rules of APA Style*”，2005）。

汇。学生有时无法理解为什么文学教授会因为转述作者的话而降低水准，而心理学教授却会由于过多的直接引用而降低水准。不幸的是，这个世界不会总那么公正，为了成功地写出一篇文章，你必须了解这些细微的差别。然而，我要强调的是，在正确引用文献时，转述他人观点和直接引用一样重要。如果你不能正确引用他人观点，那就相当于“抄袭”。我们普遍意义上的抄袭就是侵占（偷窃）他人的用语、想法和观点。不论是在学术机构还是专业机构中，抄袭都会造成很严重的后果。抄袭这一内容的详细讨论见第5章。

小标题

APA格式中报告需要有一定的顺序结构（下一节将会具体介绍）。心理学者们会使用一系列小标题来引入新的主题。使用小标题后，在引入新的主题时就可以省略一些过渡性的话语，同时也使得研究报告显得简练。在人文学科的文献中很少使用小标题。研究报告的结构性强调的是“故事图式（story schema）”，也就是说，通过一致的表述结构，研究者将他们的研究呈现给我们，通过这种方式作者与读者进行沟通，读者就可以对接下来的内容形成某种期待。标准化的结构降低了对用语的依赖，也许会削弱文字的趣味性，但保证了内容的一致与简练。

脚注

心理学家们很少使用脚注，特别是所谓的离题脚注（discursive footnotes）①。心理学家会通过脚注为所陈述的观点编辑或增添一些信息，这些信息可能是很多读者不感兴趣的内容。另一方面，历史学家们使用脚注的数量是心理学家的4～5倍，文学评论家是心理学家的2倍（Madigan et al，1995）。通过这些脚注，历史学家和文学评论家们就可以建立平行文本，通过平行文本，可以同时在多个水平上对同一问题展开讨论。心理学家却认为，这些脚注会使得读者分心，而无法将注意力集中于清晰、简练、线性格式的行文中。

分歧

当历史学家和文学评论家们无法与同仁达成共识时，他们就会直言快语地指出彼此间的不同，有时候这种分歧就造成了彼此的人身攻击。持一派观点的作者可能会指控另一方“无知”“考虑问题不全面”，甚至是“故意曲解原意”。而心理学家们却力求在分歧面前避免人身攻击。他们会通过数据、方法和理论对分歧进行说明，而不是直接指责其他的研究者。礼貌地对待其他研究者，这可能源于一个共识，即我们都是在为科学知识体系的构成添砖加瓦。无法达成一致也许会拖延知识体系的建构，不利于科学知识体系的稳固形成。而弱化个人的观点，强调数据和理论，则有利于知识体系的形成。

模糊的结论

在学术文章中，心理学家更多的是使用一些模糊限制词。与历史学家和文学评

---

① 为了强调这一点，我将使用脚注但不是离题脚注。字典里对“discursive”的解释是“离题的，没有中心的，东拉西扯的”。离题脚注就是指偏离讨论主题的脚注。

论家相比，心理学家使用模糊限制词的数量是他们的 10 倍还多（Madigan et al, 1995）。这是一些典型的模糊限制词和词组：is consistent with, lends support to, may be considered, may be related to（与……一致，支持……观点，可能被认为，可能与……有关）。当开始下结论时，心理学家们会使用很多上述模糊限制词。很多情况下，研究者保护的不是数据，而是理论。正如我们第 3 章所讨论过的那样，理论实际上不可能完全被证实，虽然可能证伪，但也十分困难。因此，即便是最具有说服力的数据，与理论的关联也极其微弱。心理学家认识到，这种弱相关可以恰当地保护他们的结论。当心理学家以一种过于武断的方式阐述结论时，其他的科研同行可能会指出，这是一种无知的行为（但会以一种礼貌的方式指出！）。

我希望对 APA 格式的探讨，将有助于你理解为何心理学与其他课程的写作风格有所不同。你仍需花一些时间充分理解不同学科在用语上存在一些细微的差异。当真正在某一领域开展实验时，特别是当你开始阅读文献或开始写作自己的文章时，才会对这一点有深入理解。

报告的组成部分

所有实验报告都应该包含某些标准的部分，这些部分应该按照一定顺序呈现。否则，我们就会像传教士说布道辞那样："首先我想说，嗯，说我想说的；嗯，然后，我说，嗯，然后，我说，嗯，说我说过的"。所有的实验报告都会按照一个标准的模式呈现，所以，我们不需要花费大量的篇幅来"陈述我们接下来要干什么"。标准化的行文组织结构不仅可以提高写作效率，还可以使得读者将注意力集中于某一部分，如方法或结果部分，从而迅速地找到所需信息。报告主要包括以下几部分（具体将在后面介绍）：

Ⅰ. 标题页
  A. 标题
  B. 作者姓名
  C. 所属单位
  D. 页首小标题
Ⅱ. 摘要
Ⅲ. 正文部分
  A. 引言
    1. 研究背景
    2. 文献综述
    3. 研究目的与假设
  B. 研究方法
    1. 被试
    2. 仪器\材料
    3. 实施程序
  C. 研究结果
    1. 结果的文字说明

2. 相关图表
3. 描述统计与推断统计

D. 讨论
1. 研究目的与结果间的关系
2. 理论与方法学的贡献
3. 未来的研究方向

Ⅳ. 参考文献
Ⅴ. 作者注(若有的话)
Ⅵ. 脚注(若有的话)
Ⅶ. 表
Ⅷ. 图标题
Ⅸ. 图

标 题

心理学核心期刊上的研究报告在出版后前两个月内约有一半会被将近200名心理学者阅读(Garvey & Griffith,1971)。读者之所以会选择某篇文献阅读,很可能就是因为它的"标题"。

大部分心理学家都会浏览期刊的封面,以查看是否有他感兴趣的最新研究。文献检索(第6章)中的关键词很多也是从标题中抽取出来的。因此,从某些意义上来讲,标题是文章中最重要的部分;如果你的标题几乎没有传递任何信息或是传递了错误信息,你就会失去大部分的读者,即便这些读者知道你所从事的研究。标题的两项基本要素是彼此矛盾的:①涵盖尽可能多的信息,②尽可能短。大部分标题都会呈现研究者感兴趣的自变量和因变量。即使变量不是很清晰,你也应该能从题目中识别大体的研究领域。*Publication Manual* 中说,标题不应该超过10到12个单词。大部分的标题应该比这一规定还短①。



一种设置标题的方法就是开始写一个比较长的标题,然后删减单词,直到你感觉不能再缩减为止。比如我们想要给"字体大小实验(the print size experiment)"(在前面的章节提到过)拟一个题目。起初,我们可能这样写:"一项关于在不同年龄段儿童中字体大小对标准段落阅读用时影响的实验(An Experiment Examining the Effect of the Size of Print on the Time to Read a Standard Paragraph for Children of Various Ages)"。现在,我们来简化一些。"一项关于……对……影响的实验(An Experiment Examining the Effect of)"就可以首先删去。因为这些单词无法向读者传递新的信息。我们也可以通过重新安排词汇顺序,而删掉大部分的介词(如 of, on, to,

① 虽然我没有相关证据,但在我看来,一般情况下,文章题目越短就越为人所知晓。这种效应的产生可能源于读者的记忆广度,或者是越好的作者越能写出简短的标题。也许我应该给这本书取名《书》。

for)。相比泛泛的描述(不同年龄段儿童,Children of Various Ages),用具体的水平来描述自变量会更加有效(8 岁与 12 岁,8-and 12-year-olds)。修改之后的标题为:"8 岁和 12 岁儿童的字体大小对阅读速度的影响(Print Size Effects on Reading Speed of 8-and 12-year-olds)"。题目涵盖了大部分的原始信息,但是更短了。

使用冒号有助于缩短标题和删减单词。如果你在这本书的结尾部分查看参考文献的题目,你会找到很多这样的例子。例如,Johnson 的一个题目:"短时记忆任务的瞳孔反应:认知过程、唤醒,还是二者都是? (Are Pupillary Responses during a Short-Term Memory Task: Cognitive Processing, Arousal, or Both?)"。还有一篇 Greenwald 的文章:"被试内设计:使用还是不使用? (With in-Subjects Designs: To Use or Not to Use)"。他可以改为:"是否该使用被试内设计(Should With in-Subjects Designs Be Used or Not)"。此时,冒号虽然不会缩短标题,但却使问题的提出更有趣味——莎士比亚式的风格。

作者姓名与所属单位

在标题之后呈现的就是作者姓名,紧接着是所属单位,即研究是在哪进行的。将研究者所在的单位列上去是十分必要的,不仅因为研究机构提供了资源,还因为它承担着维护基本职业道德、照顾好被试的责任。当有多位作者时,只列上那些对本研究有较大科学贡献的作者。那些只是协助收集、分析数据的人应该在作者标注中被感谢,而不是列为作者。一般来说,那些对研究负主要责任的作者会撰写研究报告,并作为第一作者。其他作者按照其对研究的贡献大小分列其后。但是,研究者贡献的大小并非总是一目了然,也会引发争执,特别是在教师—学生协作的研究中。避免这一争议的最好办法,就是在实验研究早期对这一问题进行商定(Fine & Kurdek,1993)。要商定的问题包括在研究中的专业贡献和非专业贡献,每个人所具有的能力,以及每个人所承担的职责。尽管随着研究的开展,商定的内容要不断修订,但提前达成一致可以避免日后出现激烈的争执。



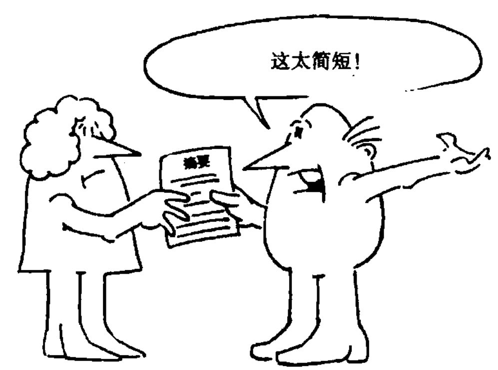
页首小标题与页眉

页首小标题是打印在标题页顶端的短标题,以便读者识别文章。页首小标题最多只能使用 50 个字符,包括字母、标点和空格在内。页眉包括标题的头几个单词和页码,页眉应位于每一页的右上角,单词位于页码的左边。一旦论文打乱了顺序,可以利用页眉重新排列。

摘 要

摘要是全文中第二重要的部分。一旦读者因为对标题感兴趣而选中了你的文献,接下来他们就会阅读在期刊、*Psychological Abstracts* 或 *PsyINFO* 中的摘要了。就像登门销售一样,标题可以让你的一只脚跨进门槛,而摘要可以邀请你进入房间。

摘要应该是全文的浓缩版。尽管终稿出来后,摘要是紧跟在标题之后的,但大部分研究者会等到全文撰写结束才开始写摘要。在摘要中,你应该介绍研究目的、变量名称,简要介绍研究方法和重要的研究结果,并讨论结果的意义。摘要不应超过 120 个单词。你必须用为数不多的词汇来表达大量的信息。你可能会发现,先写出较长篇幅的草稿,而后删减不必要的信息和语言依然是一个整理摘要的好办法,但要确保摘要的可阅读性和句子的完整性(这点和标题有所不同)。如果摘要在首次删减后仍然过长,你就要选择那些相对而言更为重要的信息,删减掉那些次要的材料,直到摘要长度符合标准为止。



一些研究者会在事后才考虑标题和摘要,他们在艰难的完成了文章正文之后,才会匆匆地撰写标题和摘要。在这一部分我想强调,题目和摘要是你实验报告中最重要的两部分,要尽力把这两部分完成好。

引 言

引言部分是用来描述相关研究的知识背景和当前发展状况的。由于它总是报告的第一部分,所以,不需要用小标题标注出来。假设读者对你研究的领域比较熟悉,你只需要提及与你的研究相关程度最高的几项实验。当你引用一项实验的时候,只需在正文部分列出作者的名字和文章的发表时间,在参考文献部分列出引用文章的全部信息就可以了①。对那些关键实验要详细地描述,以确保为你实验的提出作好准备。好的引言应该是微型的文献综述,读过引言之后,读者会知道接下来应该做什么实验——即你的实验内容。

在文献探讨之后,必须阐述你的实验目的。该段落应该详细指明本研究中自变量与因变量之间的关系。例如:“本研究的目的就是为了识别字体大小对阅读速度

---

① 如何引用一个实验?如果是一个作者,只要给出作者名字和文章发表时间“Jones(1967) found...”或者“It was found(Jones, 1967)...”。如果是两个作者,使用两个姓[Jones and Smith (1971) found...]或者“It was found(Jones & Smith, 1967) that...”。如果作者多于两个,第一次引用时列出所有作者的姓名,随后的引用只要用第一作者后面加上“et al”即可,如:“Johnson et al. (1972) also found”。

的影响效果在8岁儿童和12岁儿童中是否一样”。如果通过文献综述和理论论据可以推断实验结果,那么,就可以用假设的形式阐述你的预期结果。你需要解释假设背后的逻辑关系,以使你结果的解释更为容易。如果你的假设不合理,那就不要浪费读者继续阅读的时间了。

方　法

到此为止,读者已经知道你为何要做这些工作。接下来,你要告诉他们你做了什么。“方法”的描述应该尽量详细,这样就能使读者知道如何重复你的实验。但是,你必须很清楚,哪些细节与你的实验结果有关。例如,在“字体大小”的实验中,尽管你必须说明纸张的大小,但不需要描述阅读文字时所在房间的确切尺寸。因为环境信息太多了,你不需要一一提及,你只要介绍那些可能会影响到实验结果的环境信息就可以了。

方法部分通常会被分为几个小部分。尽管有时会由于实验需要进行增添,但标准的格式会有以下三个小部分。

被　试

被试部分应该详细描述被试的相关信息①。他们是学生、飞行员还是儿童?性别是什么?数量是多少?你是如何选择了他们?(他们是志愿者吗?他们符合分类标准吗?他们有报酬吗?)在描述这些信息时,你要保证任何背景的读者都能读懂(“被试是一群来自PSY 204的学生”,PSY 204是什么意思呢?)。如果受试者是动物,就要报告他们的属、种和供应者,饲养条件,以及年龄和性征。如果你要删除任何被试的数据,你应该指出这样做的依据。

仪器与材料

仪器与材料部分应该介绍实验中使用的仪器设备和材料。对于标准的心理学仪器,仅需列出通用名、制造商和型号就行了(“使用了科研用双通道800-F型速视仪”)。如果使用了特别定制的仪器就应该进行详细描述,以便于读者可以仿制(“幻灯片投放在一个15厘米高,20厘米宽的树脂玻璃板上,垂直放置于距被试30厘米处”)②。每次做实验时都要对所有的数据进行记录。实验结束后对这些数据进行重建是十分困难的,有时候甚至是不可能的。

程　序

在程序部分,作者必须详细介绍被试在实验期间都做了什么。撰写这一部分时,想象那些对实验一无所知的被试刚走进实验室的情景。从那一刻开始都发生了什么?指导语是什么?通常主试都会向被试详细解释这些内容,除非被试也是实验

---

① 1974年以前的文章,你会看到subject和experimenter分别用S和E代替。这些缩写法是不再被接受了。实际上,像“subject”这样的词也逐渐被更加具体的词所代替,比如“students, children或rats”。或者用通用词“participants, respondents或individuals”。

② 所有的度量都应用公制单位。如果工具上写的不是公制单位,那么,报告时应该在括号中注明换算成公制单位的数据,如“The panel was 3 ft (0.91m) in width”.

控制的一部分。测试的具体内容是什么、测试是按何种顺序进行的、如何进行时间控制？实验运用的是随机化法还是抵消平衡法？要详细描述都测评了哪些变量，怎样测评及记录的？为什么你要使用这一实验程序？

"程序"是报告中最难写好的部分之一。因为当你撰写报告时，你已经对自己的实验程序了如指掌了，此时，实验程序对你而言简单明了。一定要找一个对你这项实验一无所知的人来阅读你的程序部分，并指出对你这部分写作的看法。而后你要纠正不足之处，将改过的内容再给另一个人看。最终将这两部分进行整合，至此，实验程序的撰写就完成了。

结 果

结果部分你应该首先报告数据。只有在说明一般性的结果或小样本实验的结果时，才需要呈现原始数据。首先，应该报告描述性统计结果①。当你报告诸如平均数或中位数这些集中量数时，还应报告像标准差这样的差异量数。如果只要报告几个结果，那就可以把原始数据都列到正文上："阶段 1,2,3 的反应时分别是 50, 362 和 391 毫秒"。然而，当所要报告的数据超过 5 个或 6 个时，就可以通过图表呈现。

当因变量很重要时，研究者可以使用表格显示主效应结果以及因变量的确切值。表格和文本应该分页打印。本章结尾部分的示例报告将介绍如何制作一个表格。具体问题可参见 *Publication Manual* 或 *Presenting your findings*: *A practical guide for creating tables*(Nicol & Pexman, 1999)。

应尽可能少的使用插图，因为插图在杂志中占据的空间比较大，与表格相比，制作起来也比较麻烦②；然而，正如第 12 章所述，插图可以很好的展示交互作用和数据的变化趋势。在大部分情况中，插图优于表格，因为与表格相比，从插图中读者可以更好的提取并记住信息。以下几点是关于绘制插图的一般性原则：

1. 标记出横轴和纵轴，列出变量的单位③。
2. 纵轴(因变量)与横轴(自变量)的比例为 3:4 或 2:3。
3. 0 点标示在纵坐标上。如果为了节省空间(例如，反应时没有在 0 到 0.3 秒之间的数据)，你可以使用双斜线将纵轴切开。
4. 未在横坐标上标示的自变量，可以使用点线符号来代表。在整个报告中这些符号要保持一致。不要用不同的颜色来划分变量，除非是在彩色图书中。
5. 不要在一个插图中呈现太多的曲线，一般不要超过 3 到 4 个。
6. 插图与文本要分开，在不同的页面出现。

以上几点原则可以帮助你把结果呈现得更加清晰明了，并在最大程度上避免了数据失真。然而，你可能会发现，为了避免数据失真，有时不得不稍微改变一下。

---

① 在众多学生所撰写的报告中，我发现结果部分的第一句通常是这样的"The effect ...p < .01"。在报告描述性统计结果之前报告推断性统计结果，就好像在球类比赛中报告"其中一个队伍赢了"一样，谁赢了？赢了多少？或者在一个实验中：作用效应的方向是怎样的？效应大小如何？

② 插图没有什么标准化的格式要求。在实验报告中，坐标图是最常用的形式。

③ 新手经常会忘记这个步骤。为了避免这个错误，切记在填入数据之前一定要先标注好坐标轴。

通常来讲，要在杂志上正式发表的插图都可以用计算机程序和打印机制作出来。如果你是自己制作插图，一定要用一部好的打印机，横线的粗细和字符的大小要恰当——即使正式出版时要对插图进行压缩。为了避免插图在压缩期间变得太小，字符和数字的变化幅度不得超过 4 个 points（例如，从 14-point 到 10-point）。

描述性统计之后，就可以报告推断统计的结果了。首先，要告知读者你使用了何种检验方法；如果这一过程不清楚，就要报告你如何操纵变量进行检验的。这些检验所得的结果都以一种标准格式呈现。例如，在一项有 10 名被试的双组 *t*-检验[①]中，所得值为 4.7，在.01 水平上显著，你可以这样报告："这两组的差异显著，$t(18)=4.7, p<.01$"[②]，在报告其他统计方法时也可以采用这种方式。首先，要列上统计符号（如果不是希腊字母就用斜体字表示），然后，依次列出自由度（用括号括出）、等号"="、检验所得值、逗号、小写的斜体 *p*，小于号"<"（如果是不显著性的结果则为大于号">"），最后，还要列上显著性水平[③]。很多杂志现在都要求，在报告标准检验的统计显著性的同时，还要报告效应值（effect size）。通过效应值，读者不仅可以了解到你所得到的样本差异是否能反映总体差异，而且可以了解这一差异是否足够大。利用推断统计所得的数值就可以很容易计算出效应值。你可以通过翻阅统计教材或询问老师来获取相关的计算公式。

在结果部分不必对结果进行解释，除非有特定要求对数据进行说明。结果部分用来阐述你都发现了什么；讨论部分将解释为何你这样阐述你的结果，这二者决不能混同起来。在有些情况下，若想更加清晰有效地呈现结果，可以将结果和讨论部分结合起来呈现。此时，所用的标题就是"结果与讨论"。

讨　论

在引言部分你已经对知识体系的构成以及接下来的研究取向进行了介绍。结果部分呈现了一个新的发现[④]。接下来你要介绍新的发现如何嵌套于原来的知识体系中以及新的知识与原有知识有何不同。因此，在讨论部分中，你要利用你的结果将原有的知识体系升级重建。

在大部分情况下，你需要在引言部分介绍相互矛盾的理论，陈述可以预测结果的假设。在讨论部分，你应该简短回顾一下这些理论和假设，并探讨一下你的结果是支持还是驳斥了这些理论和假设。如果不止一项理论或假设都可用来解释结果，那么，你可以指出未来的研究方向。

讨论也是使得你的结果合理化的部分，如有必要，还要推断意外结果出现的原因（你只要简短地论述你的推测，就事论事即可）。但你无需解释那些统计效果不显著的结果，这样会浪费读者的时间。只有在极少数情况下，阴性结果才可以被理解为不是由于随机因素引起的。

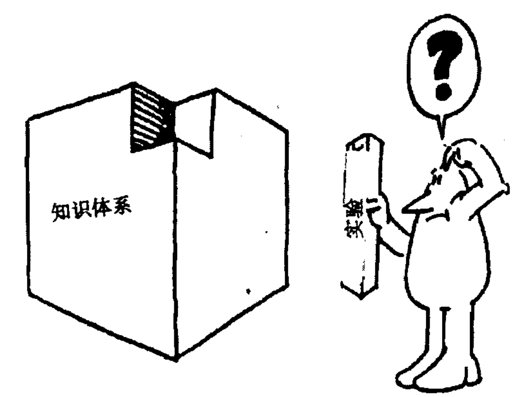
---

① *t* 检验是对两个样本的平均数之间差异的检验。通过计算得出相应的值，与界值表对比，你可以判断差异所达到的显著性水平（如显著性水平在.01，则 $p<.01$）。

② 括号中的数字是 *t* 检验的自由度。大多数的统计检验都有自由度这一项，可以是 1 个数字，也可以是 2 个数字。当你学习相应的统计学方法时，也会学习如何计算自由度。

③ 附录 A 中，在每个统计学方法例子的最后，我会列出在论文中呈现统计结果的正确格式。

④ 或者，有的时候，你的实验结果可能会否定现有知识体系中的某个部分。



值得注意的是，如果你从事的是应用性工作，那么，你就应该指出结果的实践意义——这些结果该如何应用、应用在何处，它们可以如何改善当前的研究程序。

最后，你可以通过讨论部分指出将来的研究方向。你可以就新的知识体系展开讨论，提议未来研究的开拓领域。

参考文献

参考文献部分应该列出正文部分所引用的文献，而且应该按照第一作者姓名的首字母顺序排列。本文的参考文献列在了本书的结尾部分，在报告范例中列出了正确的格式，并且有很多实用的例子。对于特殊的参考文献，应该参考 *Publication Manual*。

减少语言偏见

大多数人没有意识到，我们使用的语言都体现出历史文化偏见。例如，按照惯例我们一般使用他或他的（*he*, *him*, *his*）作为一般代名词，而不是用他或她，他的或她的（*he or she*, *him or her*, *his or hers*），因为前者在表述上更简洁；另一个原因是历史性的，即被提及的那个人一般来说就是男性。男性从事的事情更有价值，而女性只是呆在家里抚养孩子——至少掌握写作的人是这么想的（阅读这些内容的人也是男性）。我没有必要向你指出现在已经时过境迁了。我们心理学界的大多数人都坦率地承认了这一改变，大部分心理学研究生以及大约四分之三的心理学本科生都是女性。我们心理学界的男性同胞应该庆幸女性朋友们没有为了寻求平衡而要求使用女性的一般代名词。

当谈到种族、年龄、残疾以及性别取向时，也会存在语言偏见。为了减少这一情况，APA *Publication Manual* 中有一部分专门介绍语言使用的指导方针。我将这些内容整理如下，相信我并未曲解原意：

1. 个体是怎样就怎样称呼。第一点指导方针至少有两层含义。第一层含义是，无论多大或多小的团体，我们都应该尽量精确地指明。例如，如果你想把女性包含在内，就不要在"他们对知识的寻求（man's search for knowledge）"这样一种叙述中，仅仅使用"男性（man）"这一词，而应该使用"男性和女性（men and women）""人类（human beings）"或者其他包容性更广泛的概念。同样，应该使用"他或她，他的或她的（he or she, him or her, his or hers）"这样的表述，代替"他"或"他的"（he, him,

his),也可以将句子换为复数形式——例如,将"问被试他是否曾……(Each participant was asked whether he had...)"变为"问被试们是否曾……(Participants were asked whether they had)"①。通过这种方式你的语言会更为精确,也没有任何群体会被排除在外。另一方面,避免使用涵盖面积太广的词汇。如果你想指的是"非裔美国人(African American)",那就不要用"非白人(nonwhite)"一词。换句话说,用词详尽就可以保证用语的精确性。

第二层含义就是将人作为人,而不是客体。在本书的其他部分我已经讨论过,应该尽量不要使用"subject"这个词,因为这种称呼显得被试像一个物体,而不是人。因此,应该尽可能详尽:儿童、学生、小白鼠、8岁、女性。若要使用更加泛化的词语,就可以使用"participants, respondents"。

2. 用人们乐意接受的方式称呼他们。我们对词汇的运用会随着时间的改变而改变,我们对某些群体的称呼也会随时间而有所改变。有时候这些称呼的变化十分迅速,以至于再过几年我这本书上罗列的最新名词也会过时。在过去50年里,我们对黑人的称呼依次经过了以下几种变化:Negro, colored, Afro-American, black, African American,现在很多人称之为 people of color。随着这本书的出版,亚裔美国人(Asian American)逐渐取代 Oriental 一词。对印第安人(Indian)应称为 American Indian 或 Native American。当提到"性取向(sexual orientation)"时[注意,"取向(orientation)"是个中性词,好于"选择(choice)",因为我们不知道是否存在"选择"问题],一般会用 gay man 或 lesbians 。不论何种情况,你都要确定在你的论文写作中,使用哪种语言比较恰当。最好的办法就是询问参与者比较喜好的命名。

3. 人们是名词,他们的特性是形容词。这一点说明,人就是人,他们不是一种属性,也不是一种状态。因此,有精神分裂症的并不是精神分裂症患者,精神分裂的只是一种疾病类型,并非一个个体。与此类似,身患残疾的人不应该被称为残疾的,年长者不应被称为年纪大的。同性恋者不是同性恋。其他形容词也适用于这一规则,如男性和女性,应该写女性参与者,而不是女性。名词的称呼就是男人和女人,高中生或更年轻的人就可以称之为男孩或女孩。顺便说一下,一定要使用平行词汇,特别是当非平行词汇将一组被试变为次级类或专属类别,如男性和妇女就是两个非平行词汇。

以上几条只是建议,而不是严格的规定。有时候,文章会因为刻板的遵循以上的指导方针而显得冗长繁复。我们不应该将无法达到以上规定作为降低科学准确性的借口。Maggio(1991)建议,你可以通过下列方法检查你的写作是否含有暗示的评价,即想象你自己是所要讨论的团体的一分子,如果你感到自己被排挤在外,那么,你就需要对原文进行修改了。

写作风格

实验报告不是文学名著,也不是舞台剧。因此,你的整体写作风格不应该妨碍

① 不要混淆单复数,比如"Each participant was told that they could..."。在演讲时,如果因为句子太长可能会出现单复数混用的情况,这时候人们不易察觉。但是在成文的论文中人们就很容易发现了。参考 Foertsch and Gernsbracher(1997)和 Madson and Hessling(2001)的文章,看看具体的错误。

思路表达的流畅性,也不应该把读者对研究的注意力过多的转移到你身上。为了满足以上几点,科研写作需要有一个标准的格式。

按照惯例,科学写作都使用第三人称的被动语态,而非第一人称的主动语态。我们一般这样陈述:"这项实验被操作用来……(This experiment was done to...)",而不是"我做这项实验是为了……(I did this experiment to...)"。尽管这样使得报告阅读起来不像是在读家书,但这种做法也削弱了文章的文采。整篇文章单调乏味,读者在阅读的时候也感到艰难晦涩。如今,在一定程度上也可以使用代词"我"——例如,"我认为……(I thought that...)",而不是"这件事被认为……(It was thought that...)"。但是,还是应该避免过度使用"我",因为这样读者就会将注意力投向你,而不是研究本身。还有一点就是,你应该使用动词的主动语态,而不是被动语态,特别是不存在代词问题的时候。例如,"以往的一份报告介绍了一种新的方法(A previous report described a new method)",而不是"在一项以往的报告中,一项新的方法曾被介绍(In a previous report, a new method was described)"①。大体原则是,词汇的使用要确保写作的生动性,而不是破坏思维的流畅性。

句子的语境会告诉你应该使用何种时态。在引言和方法部分的大部分语句应该使用过去时"Boles(1972)曾报告……"[Boles(1972) reported...],还有"学生回忆了单词……"(The students recalled the words...)。此外,即使是实验已经结束,对理论和结果的阐述都要用现在时。也就是说,对当前知识体系的阐述应该用现在时:"这些数据支持了遗忘的干扰理论(These data support an interference theory of forgetting)"。

最后,科研写作要简明扼要。因为时间和空间有限,我们不能说空话。例如,我在这本书中所用的写作风格就不适用于科研写作②。我有意使用了一些附加的内容。因为我不仅仅想传递信息,除了交流以外,我还想向读者证实一些东西,引导读者接受一些东西、转化读者的一些观点。在科研写作中,你可以假设读者已经接受了这些观点,你唯一要做的就是与读者进行交流。

新手在科研写作中最常遇到的一个问题就是懒惰。当然,研究者并非真正的懒惰,因为懒惰的人是不会做实验的,我是指他们的写作风格比较懒惰。在你写作的过程中,最重要的是要写到铅笔秃头了为止,最重要的键是删除键。在第一遍写作时,很少有人可以写出简洁流畅的报告。大部分优秀的科学研究者在选出最好的一稿之前,都要几易其稿。每一个词汇都必须精确的体现出你要表达的意思,句子间也要衔接自然。而这是一项很艰难的工作!

写报告时,多数人都会先拟一个草稿,并不断修改,从而达到最佳水平。为了呈现最佳的报告,一般都要修改两到三遍,因为将整篇内容推翻从来要比不断修改达到最佳水平容易得多。一旦报告的撰写已经使你满意,接下来就应该将这份报告拿给其他人阅读。这些人当中应该至少有一名是不熟悉你实验内容的。你可能对你的实验内容过于熟悉,以至于无法判断该如何描述它,也会使得你对报告中遗漏的

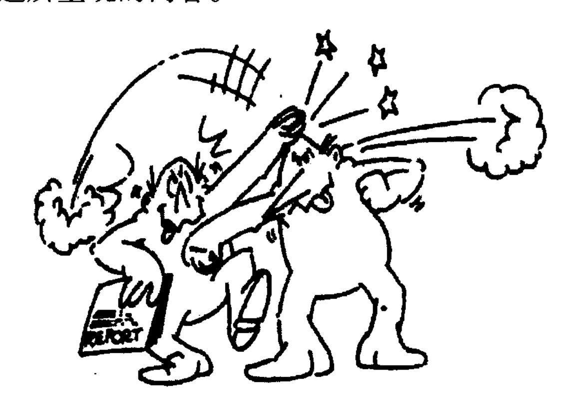
---

① 有些作者可能会就此提出质疑:报告不会描述,人才会描述。我想每个人都有不同的写作风格。在这个例子中,我觉得牺牲一点准确性,增加一点趣味性,未尝不可。

② 如果我用科研写作风格撰写这本书,你、我还有出版商都会感到无聊。但我妈妈不会,因为她爱我。

问题熟视无睹。因此,不了解情况的人会是一个很好的细节补充核查者①。

还应该让熟悉你这项研究的人阅读你的这份报告,通过这种方法他可以告诉你报告上的内容和实际的操作是否相匹配。这种人可以称之为错误核查者。他们可以告诉你该如何改进所呈现的内容。



敌人是最好的批评者

在征求完这些人的意见之后,你就可以着手写作这一报告的最终版了。在递交报告之前,你要用心打字、印刷、检查拼写及校对。

你也许会认为,遵循了以上的规则,报告的可读性会更强;但有人会认为,其他写作规则的效果更好。写作是一门艺术,适用于一位作者的规则未必适用于另一位作者。然而,有一点适用于任何一种规则,也非常重要,即报告是你研究的最终结果,你在撰写报告过程中所投入的努力应该不啻于你在研究中投入的努力。

常犯的十大错误

我批阅过大量的学生报告。即使我已经教给了他们 APA 格式,他们还是经常犯错。其中一些错误出现的频率很高,因此,我认为,有必要列出最常犯的 10 个错误。具体如下(当然,不可否认,还可能会出现其他错误)。

APA 格式中学生最常犯的 10 个错误

10b. 当没有时间词时,使用“since”,而不是“because”或“as”。

10a. 在圆括号内使用“and”或在文本中使用“&”。

9. 引用文献时,在“et al.”中 et 的后面加句号(et. al.)。

8. 将被试称为“subjects”。

7. 参考文献是按卷分页码的,依然给出期号。

6. 当呈现统计结果时,小于号(<)或大于号(>)用反了。

5. 小标题的层次设置错误。

4. 未将所有引用文献列入参考文献部分。

---

① 在一个同事的追悼会上,我的朋友告诉我“我会想念他的,他是我最好的‘敌人’之一。我现在不知道应该把我的文章发给谁看了”。通常,一个人能严厉地审阅你的文章,而不怕破坏你们的关系,这是最好的阅稿者。朋友通常会不好意思过于挑剔。

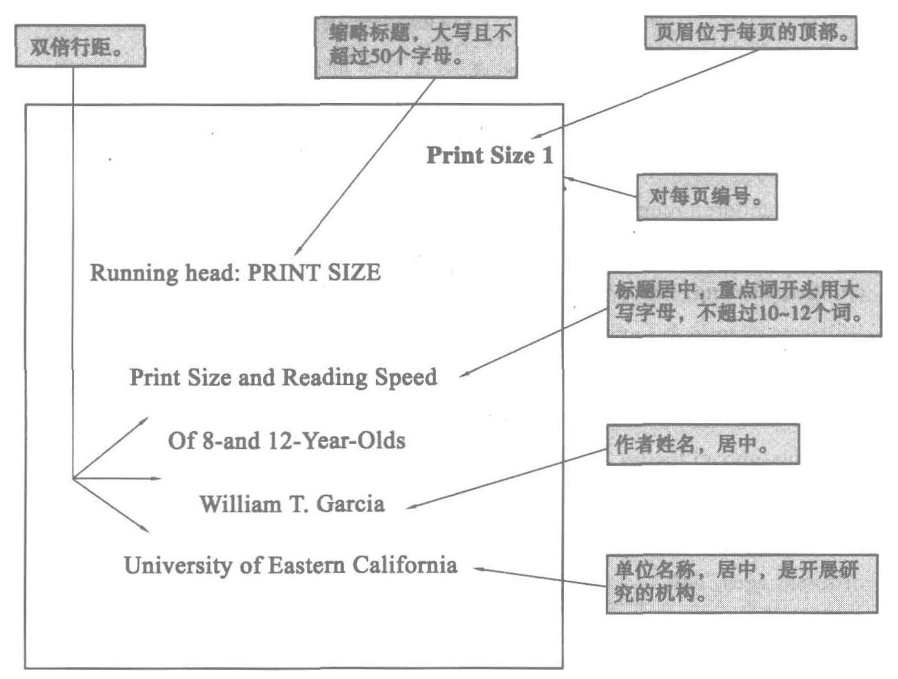
3. 写出"data is"的表达(data 是复数词,应该为"data are")。
2. 误将"male"和"female"这样的形容词作为名词使用。
1. 过去从事的研究未使用过去式,当前结果和理论未使用现在时。

报告实例

请不要太在意报告实例的具体内容,因为它不仅是虚构的,而且其写作风格不佳。我只是希望通过这个简短的报告,尽可能多地说明 APA 格式的应用实例。在实例中需要注意的地方,我会用箭头标注,并在空白处作格式规则的简短说明。详细规则还得阅读 *Publication Manual*。

你的导师也许不要求你遵循某些规则。例如,如果一篇报告不是为了在杂志上发表,多数情况下,我让学生把图表穿插在正文里。这样读者会更容易理解。你的导师或许也有这样的喜好。

如想了解 APA 格式的更多细节,你可以参考 *Publication Manual of the American Psychological Association*(第 5 版, 2001)。如果你在写作方面需要更多指导,以下是一些可参阅的资源。Fred Pyrczak 和 Randal R. Bruce(2000)出版的 *Writing Empirical Research Reports: A Basic Guide for Students of the Social and Behavioral Science*,这是一本很好的参考书。如果你对 APA 格式的运用需要更多的指导,请阅读美国心理协会出版的 *Mastering APA Style: Student's Workbook and Training Guide*(Gelfand, Walker & APA,2002)。一些软件产品能帮助你更容易地掌握 APA 格式,比如在 http://www.apastyle.org 下载 *APA-Style Helper* 5.0,在 Reference Point Software.com 下载 *Reference Point Software's Template for APA Style*。



摘要不缩进。

Print Size 2

Abstract

The time it took 8-and12-year-old children to read paragraphs was measured as a function of print size. Paragraphs were typed in either 12-or 10-point print. Results indicated that the size of print is an important dimension for the younger children, with the larger print being read 32% faster than the smaller print. Print size had no effect on reading speed for the older children. The results are interpreted in the context of the stimulus-impact theory of reading.

双倍行距。

不超过120个字。

不要求右对齐，单词不断开。

所有页边距为1英寸。

段前缩进5到7个空。

再次呈现标题，居中。

Print Size 3

Print Size and Reading Speed
of 8-and 12-year-olds

Many researchers have speculated that younger children require larger objects to have a given visual impact compared to older children(Darling, 1990; Millar & Hoaks, 1989; Williams, Garcia & Levitz, 1992). In their classic review article Millar and Hoaks(1989) specified the dynamics of this statements and called it stimulus-impact theory.

Williams et al.(1992) later tested the theory by measuring pupillary response in children at three different ages. The 3-year-old children required a 375-ms longer exposure to produce the same size response as the 7-year-old children.

双倍行距。

引用文献和理论时要附上作者姓名，按字母顺序排列，用括号括起来。

在括号中使用“&”。

在第一次提及理论时至多列出6位作者的姓名。

在第一次引用后，如有3位以上的作者，用（“et al.”）。

数值后的单位缩写，如毫秒用“ms”表示，且“ms”后不用句号。

除摘要外，数字0~9都用文字表示。

对时间、日期、年龄和带单位的数字，用阿拉伯数字表示。

在句子中使用“and”

即将出版的文献用"in press"来标注。

用国际通用计量单位。

Print Size 4

10或以上的数字用阿拉伯数字表示。

Grant(in press) also used pupil size to compare the effect of 10-cm and 20-cm disks on 48 children of various ages. Disk size had little effect on children over 10 years old but had an effect on younger children.

In the research reported here I attempted to apply stimulus-impact theory to a reading task. On the basis of this theory I predicted that print size would have little effect on reading speed of 12-year-old children but would affect 8-year-olds.

Method

Participants

Thirty 8-year-old children from an elementary school, 15 girls and 15 boys, served as

5个层次的标题的版式如下：
1.所有字母大写，居中。
2.单词首字母大写，居中。
3.单词首字母大写，斜体，居中。
4.单词首字母大写，斜体，左对齐。
5.小写，斜体，缩进，以句号结尾。
如果只有两个水平，使用第2和第4条。

尽量不要以数字作为句子的开头，如要用，用文字表示数字。

Print Size 5

Participants in one group.Their principal and a parent had given permission for them to serve. The 17 boys and 13 girls in the 12-year-old group were paid $ 3.5/hr for participation.

被试信息包括人数、性别、年龄、抽样方法。

Apparatus

A slide projector back-projected the paragraphs onto a 30-cm × 40-cm frosted-glass screen. A stopwatch was used to measure reading time.

详细介绍特定仪器。

简单介绍一般仪器。

Procedure

The children were individually tested in 20-min sessions.After being seated in front of the screen, they were told that a paragraph would appear on the screen.They were to read this paragraph carefully, taking as much time as

Print Size 6

necessary to understand the material. After reading a paragraph,each child was asked three questions, having single-word answers, about the contents of the paragraph. After the questions were answered, another paragraph was presented until each child had read three paragraphs.

Each paragraph had been previously tested for readability and was at or below an 8-year age level. The questions had been found to be a good measure of comprehension.

The experimenter manually timed the reading latency for each trial using a stopwatch. Scores were obtained for each of the three trials in

步骤用过去时。

十以内的数字用文字表示。

表示时间、年龄、分数或百分比都用数字表示。

Print Size 7

each session.Thus, the experimental design was a 2 × 2 × 3 factorial design having age at two levels (8-and 12-year-olds),print size at two levels(10- and 12-point),and trials at three levels.

Results

Mean reading times for each age group and print size are show in Figure 1.An analysis of variance computed on reading times indicated that the main effect of age was statistically significant, $F(1,58)= 26.73, p<.01$. the main effect of print size was not significant,$F(1,58)= 0.87, p<.05$. however, the Age × Print Size interaction was significant, $F(2,116)= 10.31, p<.01$.

如果是真实的报告，此处数据点太少，不适宜作图或表。

在文中标明图标题。

注意怎样报告统计结果。

若数字小于1时，小数点之前要加0；若数字不可能大于1的情况，则不加0（如相关系数、显著性水平）。

以A × B的形式表示交互作用时，首字母需大写。

主效应无需大写。

报告均值时，同时要报告离散度，如标准差。

Print Size 8

Mean reading times and SDs for each of the three trials are shown in Tables 1. The main effect of trials failed to reach significance, F(2,24)=1.53, p>.05.

此处用这个表仅仅为了演示表的用法。在真实的报告中，不会用表和图来展示同一组数据。

注意<、>号的方向；p>意味着不显著，p<意味着显著。

推断统计结果（如t检验和F检验）应保留小数点后2位。

Discussion

The present data are entirely consistent with stimulus-impact theory. No difference in reading time as a function of print size was found for older children. However, for younger children a print size difference caused a significant difference in reading time, An interpretation of these data within the framework of stimulus-impact theory is that even the smaller print size

数据和解释应该使用现在时。

Data是一个复数词。

Print Size 9

had maximum visual impact on older children. The younger children required a larger-sized print in order to perform at a high level. As Miller and Hoaks stated in their 1989 article, "High-impact stimuli are necessary for maximal performance in younger children"(p. 346).

在引用40词以下的引用文时用引号；长于40词的引用文不用引号，但要另起一行并缩进排版。

The implication of these results is obvious for publishers of children's reading material.However, before recommendations can be presented to these publishers, additional research is needed to compare reading times for many additional print sizes and for children at many age levels.

第一行左边对齐，其后几行要缩排。

以作者名的首字母顺序排列。

Print Size 10

参考文献中的每个条目都要在报告中被引证过，同样，报告中引证过的条目都要在参考文献部分列出。

References

Darling, D. T.(1990). Internal consistency in stimulus-impact theory. *Journal of Child Behavior* 26, 58-63.

Grant, U. T. (in press). Pupillary response to disks. *Sensation & Perception.*

Millar, J. R., & Hoaks, A. R. (1989).Stimulus-impact theory; A developmental theory of perception, *Childhood Perception and Cognition* 7, 278-295.

Williams, E. T., Garcia, W. T., & Levitz, G. W. (1992). A review of size effects. *Behavioral Review*,21,326-354.

作者：姓氏加名的首字母。

出版的年份。

篇名：首词的首字母大写。

刊名：重要词的首字母大写。

起止页码。

期号：刊名（斜体）和期号间用逗号隔开。

有多个作者时用“&”。

你可以通过本书最后的参考文献部分，了解书、杂志和其他参考资料的具体书写格式。

Print Size 11

Author Note

William T. Garcia, Department of Psychology (now at the Center for Child Development, Westbrook University).

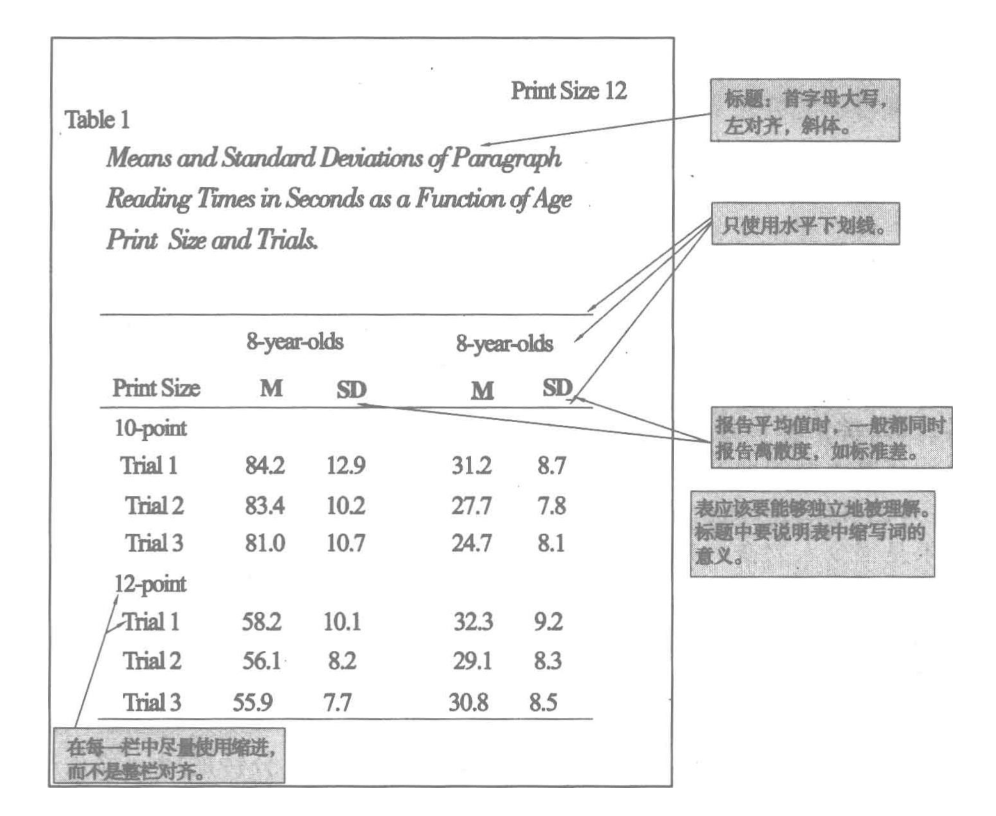
I would like to thank Nancy Wells for her help in data collection. This experiment was reported at the Northwestern Psychological Society meeting in Madison, Washington, May 15, 2001.

Correspondence concerning this article should be sent to William T. Garcia, Center for Child Development, Box 4546,Westbrook University, Monroe, Washington 12342,or by electronic mail to garcia.ccd.wu.edu.

第一段介绍作者和所属研究单位。

第二段感谢资助单位和对研究有所帮助的人，并交代此研究的前期报告情况。

第三段给出联系方式，包括现在的通讯地址。



Print Size 12

Table 1

Means and Standard Deviations of Paragraph Reading Times in Seconds as a Function of Age Print Size and Trials.

	8-year-olds		8-year-olds	
Print Size	M	SD	M	SD
10-point				
Trial 1	84.2	12.9	31.2	8.7
Trial 2	83.4	10.2	27.7	7.8
Trial 3	81.0	10.7	24.7	8.1
12-point				
Trial 1	58.2	10.1	32.3	9.2
Trial 2	56.1	8.2	29.1	8.3
Trial 3	55.9	7.7	30.8	8.5

所有图的说明打在单独的一页。

Print Size 13

Figure Caption

Figure 1. Mean reading time in seconds as a function of print size and age.

在图背后接近上边缘处写“Top”、页首小标题及图编号。

有标题的表要单独列出以使读者明白其意义。标题中要解释图中使用的所有简写的意义。

纵坐标的长度是横坐标的2/3到3/4。

图例应标示在图的边线内。

使用容易分辨的标志，如环和实心三角形。

90 75 60 45 30 15 0

Reading time (in seconds)

8-year-olds

12-year-olds

10-point 12-point

Print Size

字母大小差别不超过4个号。

标出坐标轴的名称及单位（可以缩写）。

图中的字母和数字必须足够大，以便在杂志缩小时它们也是清晰的。

会议论文

口头报告

虽然报告研究的主要形式是书面论文，但多数研究者们（包括许多学生）也会在专业会议上口头报告他们的结果。学生们有越来越多的机会进行口头报告。例如，去年，我所在大学的学生既能在心理学院，又能在全校的学术会议上以墙报的形式介绍其工作成果，这两个会议都是由学生自己组织举办。他们也有机会参加本科生学术会议（学生来自各个州）或州立心理协会主办的会议，并递交论文。一些学生甚至能在更大型的地方性或全国性学术会议上提交论文或墙报，他们可出现在学生分会场，也可作为一名合作者出现在主会场上。许多学生第一次参会时，会觉得展示结果是很令人兴奋的事，因为这既可以向其他人介绍自己的工作，同时也可以了解相关领域更广阔的科学工作。然而，提交论文或墙报可能会是一个令人害怕的经历，除了读有关演讲的课程外，多数学生都没有受过多少相关培训。以下是关于口头报告的速成课。

会议论文需要包括哪些内容呢？虽然这一问题的答案因会议不同而千差万别，但通常来说，你都要站在20—100位同行面前，以10—20分钟的时间报告你的工作，并在最后留一些时间来回答问题。一般情况下，你需要准备一些可视的材料，这些材料如今大多用电脑如PowerPoint形式呈现，某些情况下也会以打印稿的形式呈现。值得注意的是，口头报告所用的打印稿与书面论文大不相同。如果你试图在会议上向你的听众读一篇APA格式的研究报告，那么，他们不把你轰下台已经是幸事了。书面的报告太长，包括了太多的细节，并且采用的不是口头对话的格式。你可能希望读你书面报告的人能够完全理解这篇报告并能总结出许多细节。我们从关于人类记忆的研究中了解到，听你报告的人也许只能总结出少数几个要点。你要保证听众能够很容易的理解和记忆这些要点。

在理想情况下，你呈现报告的风格应该是会话式的/交流式的。我清楚地记得我在会议上呈现的第一篇论文。那一次，我演讲的对象均来自于国家级研究机构，而当时我进入这一研究领域才刚刚5个月。我当时非常害怕。为了能脱稿，我下了很大的功夫记住了我将要报告的内容。轮到我的时候，我双腿发抖，但我很流畅地报告了大量内容。我对自己的表现十分满意。但在随后的招待会上，我的导师询问会议的主办者："David今天在会议上的报告，表现得不错吧？"主办者回答说："不。他像是在读论文，而在这个会上不需要这样做。"我在很长一段时间内深受这次对话的打击，但这也是我得到的最好的建议。参加交流活动的研究者们都知道，交流性的演讲与读文章大不相同。在交流性的演讲中，我们会根据心理活动来调整速度，停下来思考下一句话，然后，迅速地表达出来。听众利用演讲者的停顿和语速变化来了解演讲者的思维模式。在读文章时，由于词与词之间没有长的停顿，词节奏更为规律，这种单调的形式很容易使听众昏昏欲睡，而无法感受到研究给人带来的兴奋。作为一名报告者，你应该十分熟悉你要报告的内容，但不要排练太多次，否则，听上去就会像在读一篇书面报告那样令人厌烦。

如果你希望以尽可能自然的方式作报告，准备材料的时间可能比你预计的要长。如果你在练习时朗读书面论文，即使读的声音很大，你也会读得越来越快。因为你已经理解了论文的内容，不会放慢速度以使听众完全理解消化。如果听众第一次听你的报告，你应该把报告时间加长 20%。理想的做法是，你可以准备一些备用的部分，根据报告的时长进行调整，或者增加内容或者省略。如果你的报告与我的类似，那么，你应大大缩减备用部分，而结束你的演讲。

设身处地多为你的听众考虑，并尽量为他们调整你的演讲内容。你的听众是学生？心理学家？你所在领域的专家？还是其他学科的科学家？在报告自己的研究时，由于对讲述的主题非常熟悉，我们很容易忽略一个事实，即并不是每个人都像我们一样对这个主题那么熟悉和感兴趣。试着回到你第一次获得研究思路的时候，这可能就是你报告的开始。

许多年前，我可能还因为报告者没有采用足够多的图表而批评他们。由于图表的制作和展示比较困难，所以，报告者提供得太少。而如今，我觉得报告者陷入了过多呈现图表的误区。使用诸如 PowerPoint 这样的软件，你可以很容易地把所有观点呈现在电脑屏幕上。我认为这样做是非常不可取的！虽然我还不至于像 Edward Tufte 那样写文章和书来表达对 PowerPoint 的不满，但我确实也认为，这一工具会导致呈现技巧变得更差①。

我认为，成功的报告者（以及老师）善于讲述有吸引力的故事。想象一下以 PowerPoint 的方式来呈现儿童的睡前故事。我觉得我找不到更好的办法来表达故事的神秘和情感。同样的道理，一旦你将自己感兴趣的研究故事转化成一系列幻灯片，许多听众将会重点快速浏览内容，而无法认真听你的讲述。我倾向于只在幻灯片上呈现少量要点，逐一呈现，逐一叙述。我觉得不使用 PowerPoint 是很合适的，尤其是对那些经验较少的报告者；你的所有内容都在屏幕上呈现了，这样就很容易照着它们念。而我的建议是，为了避免这一点，去试着讲述一个故事，仅仅穿插着利用 PowerPoint 来强调要点。

在听取大量会议报告之后，我认为，报告者所犯的最大错误是对实验程序一语带过，而迅速到达所谓最重要的部分——结果。除非你的过程非常清楚，否则，结果并不重要。我发现，帮助听众全面理解过程的最佳方法是，在可能的前提下让他们亲身参与。如果在我的实验中，我向被试显示一系列可视材料并记录他们的反应，那么，在我的报告中，我会给听众提供一份简要的说明（呈现实验中的可视材料），并要求他们作出相对应的反应。呈现几个代表性的试验，比只利用言语解释实验过程能让观众更好地记住实验过程，花费的时间也更少。心理学家知道，比起被告之如何做，人在做中学的效果更好。我们要好好运用这个规律。

在报告论文时，你要有一个大体的规划。如果听众真得很想了解你平衡处理的步骤、检验的统计显著性水平或其他类似细节，他们会在事后问你或者向你索要一份书面形式的论文。而大多数听众不会记得这些细节。一般情况下，报告结果的最佳方式是用电脑呈现图表。如果条件允许，你应该在面前放一个电脑来显示在你身

---

① 如想了解 Tufte 的评论，请见 http://www.wired.com/wired/archive/11.09/ppt2.html. 如想了解 Gettysburg 对 PowerPoint 呈现的看法，请见 http://www.norvig.com/Gettysburg.

后的屏幕上放映的内容。这样的安排将使你在介绍幻灯片内容时能处在话筒旁并一直面对观众。如果你在介绍幻灯片时,用了像指示棒之类的指示物,以强调幻灯片放映时的某些重要内容,请注意在转向屏幕时加大说话的声音。有个更好的解决办法就是,使用可以通过操作电脑来指示幻灯片的程序,这样就不必在介绍时转过身去。相较于小册子,我更喜欢用幻灯片,因为后者有利于在展示中指示一些内容。在听众的小册子上,你无法实时指出展示的内容。使用小册子还有一些问题,包括分发小册子需要花费时间以及在报告时失去对呈现内容的控制,因为听众会在你报告之前就阅读小册子以获取信息(我只要有小册子在手,便不会听报告者的介绍而快速查看结果部分)——更不用说这样做的花费、造成的会场混乱和对报告结构的破坏。

在播放报告时,要保证你所处的位置不会挡住观众看屏幕的视线。当你展示一个图表时,要清楚你的观众是第一次看到这个图表。报告者经常以很短的时间呈现一个图表,然后,又马上得出一个结论:"你们可以清楚地看到,这个结果支持了我们的假设。"而作为一名观众,我在心里这样说:"等一等,每一个坐标轴代表什么?图上的升高和降低哪个表示结果变好?实线和虚线各代表什么情况?在不能支持结果时,图形应该是什么样的?能支持假设的图形是怎样的?"观众不可能有机会问所有这些问题。所以,你应该在呈现图表后,暂停一下……解释每一个坐标轴分别代表什么,解释线或柱各代表什么,并提示观众在图表的哪个部分可以获得重要信息,就像你自己看图表时的步骤一样。最后,要保证图表足够大,与背景的区别足够明显,以使坐在房间最后面的观众都可以看清楚。因此,书面报告中的图形可能要重新编绘。在你对图表完全满意之前,找一个与作报告的环境类似的房间进行演练。每次播放时,都要走到房间的最后,观察幻灯片上每个细节是否都能看清楚。在电脑显示器上看起来很协调的颜色和背景,往往在用屏幕放映时变得不一样,材料很难被看清楚。幸运的是,随着现代文字编排软件的发展,画一个漂亮的图是一件非常容易的事,但在我所参加的许多会议上,我仍然看到了许多不合格的图表。

在你报告的最后应该列出你的结论。这是你最后一次向你的听众传递可以让他们留下印象的信息。这一步十分有意义的。观众可能记住的结论是 3 到 5 个。在给出结论后,你要准备结束报告了。"嗯……这差不多就是我想要说的"不是一个令人印象深刻的叙述。如果你说"谢谢你们的关注"或"如果时间允许,我乐意回答大家提出的问题"。这将很好的提醒观众,报告已经完成并让他们明白你精心准备了你报告的结束部分。

你终于完成了报告,并准备坐下、深呼吸、放松。终于解脱了!但你论文报告小组的主席却说:"我们还有一些问题要问。"当然,你对这些问题没有丝毫准备,你已经在表达十分清楚的报告中回答了几乎所有可能被问的问题。然后,观众中有人喋喋不休地问道:"我不明白你为什么宣称你的结果支持了 Landon 的理论。难道 Wagner 去年提出的衰减理论不能预测你的结果吗?"当然,你从未听说过 Wagner。

你该怎样回答？我不能给你固定的答案①。我的建议是你要尽自己最大的努力来设想你的问题。在你自己考虑完了所有你可以考虑的问题后，你可以请别人再给你提些问题。其实，准备论文呈现的最好方式是在一群同行面前做一个预讲，这些同行可以是你的同学，也可以是你学院中的其他学生和工作人员。尽量鼓励他们向你多提有难度的问题。试着当面回答他们的问题，并在有时间的时候再想想怎样以更好的方式回答这些问题。这其中的一些问题很可能在会议上出现。所以，一定要作好准备。

海　报

如今大多数的会议都有口头报告和海报两个环节。想象一下，在一个大房间中摆放着一排排独立的展板，作者站在钉着自己海报的展板前，人们在海报前驻足，有的只是读读标题，而有的会与作者交谈。特别的是，报告者有大约 1 小时的时间可以站在其海报前解释其研究，并与不断变化的观众一起讨论他们所感兴趣的内容。

相较于论文，海报的优点是，你有机会进行真实的互动对话，互动的对象也通常是对你的研究感兴趣的人。对于那些实验设计简单、结果明确到可以简单的用几个图表进行描述的研究来说，这一形式的效果特别好。海报的缺点是，当你花 2 分钟向第一位观众解释清楚了一个问题时，可能下一位观众问了同样的问题。如果你重复解释这个问题，第一位观众会觉得无趣，而如果不解释这个问题，第二位观众又无法很好地理解你接下来要讲的内容。这样的情况往往在海报环节一再上演，到最后你会发现你甚至不能够向任一位观众从头至尾的解释完你的研究。当你的研究很复杂，或采用了不常用的复杂方法，或涉及过于琐碎及不为人熟知的理论时，这一缺点显得尤为突出。在这种情况下，你可能就没有时间来详细讲述你做研究的理由。

你需要尝试着做一些小型报告来为海报环节作准备。用不长于 1 分钟的时间来概括你全部的研究工作，这是为只想对你的研究了解一个大概的观众准备的。你也要整理一个长约几分钟的“半小型”的陈述，这是为对你的研究表现出相当多兴趣的人准备的。你还要准备一个比较全面的陈述，以此与少数的同领域的研究者们进行交流。你还应该投入更多的精力，以使你的海报做得简单明了，不言自明。你的海报能越好的解释你所做的工作，你就能更容易的将时间花在与观众的互动交流中，而不是一再重复解释基本的问题。那么，海报应该包括哪些内容呢？

当你接到展示海报的邀请后，主办方应该会给你一些关于布置的细节要求。一般情况下，你会获得一块 4 × 8 英尺的展板和一些图钉（你可以以任意方式来钉你的海报）。你要将你的材料添加到这块展板上。总而言之，不要简单地将你的论文裁剪，然后贴在展板上！这样的材料字体太小，细节太多，没人愿意花时间来读它。要知道你的观众只会花少量的时间来试图理解你所做的研究工作。你要在这少量的

---

① 我认为在这种情况下，最重要的是诚实，但我也见过报告者这样回避尴尬的局面。可以分为以下几类：(1)时间太紧没法现场回答——“您的看法不错，但这个问题现在讨论起来太复杂。我们下次有机会再讨论行吗？”“我考虑过这个问题，但最后由于某些原因放弃了，这些原因过于琐碎，在会上时间有限，我们不太可能深入讨论。”“我认为，他的理论不能在我实验的条件下直接使用，但我愿意以后与你就这个问题进行讨论。”(2)告诉我更多——“您认为他的理论在什么方面适用于我的结果？”“您能讲得更详细一些吗？”“我想听听你对那个问题的看法。”(3)这不是我的错——“我的合作者会很愿意回答这个问题。”

时间内尽可能地向他们传递更多的信息。在这种情况下,一图胜千言。

图 13-1 是一个海报布置的样本。关于论文用纸型号的信息会先行给出。你会发现大部分的材料都是图表。总的策略是,尽可能少用文字材料。如果你确实需要添加文字,那就试着把它们放在图表说明部分。第二个原则是,信息块要从左上角向右下角逐列排版,而不像英文文章中从左至右逐行排版。如果你的海报需要来回观看,那么,观众的人数将大大减少。将信息逐列排版,我们就能解决这个问题。像图 13-1 那样给每个信息块标注数字,或者使用箭头,对于观众都是一个很好的指引。

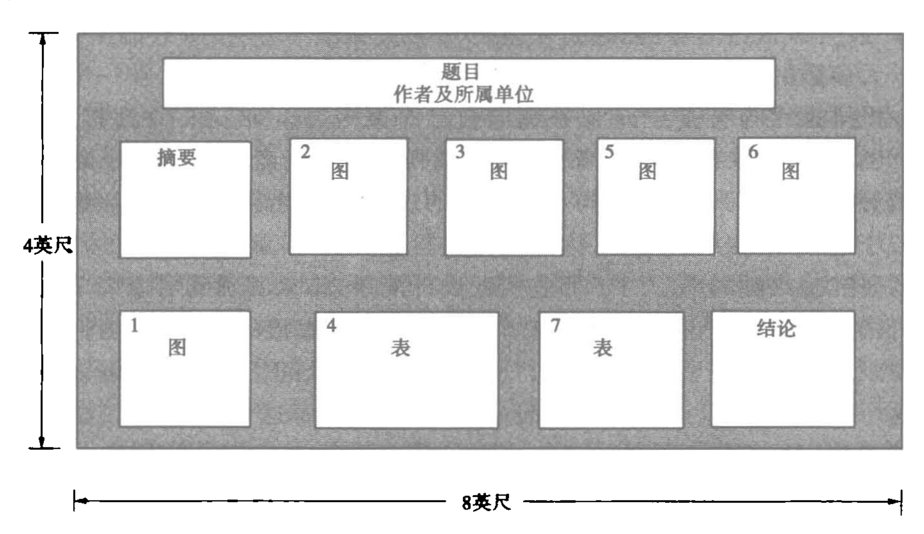


图 13-1 海报布置的样本。标题和作者的字母至少要 1 英寸高,其他印字至少 0.33 英寸高。信息要从左上到右下逐列排版。图表部分尽量做大,文字部分尽量做小。

在展板的最上部应设置一个标题以及作者名单。这些部分至少要达到 1 英寸(72 points)高的规格。要知道海报要能被 1 米或更远距离的观众看到。标题要足够大,以使观众能越过与作者交谈的许多观众的头顶看到它。许多观众只是看看海报的标题,如果不感兴趣就继续往前走了。在作者的名字下边要附上其单位,如学校或学院的名称。这些部分以及接下来的全部内容可以用较小的字体,但高度仍要在 1/3 英寸(24 points)以上。图、表、画和注释都与幻灯片上你在报告时用的一样,采用简单的粗线。为了检验材料的可读性,我所在的一个组织主张用"脚尖测试"(Human Factors and Ergonomics Society,1995)。把你即将展示的材料放在地上,邀请一两位朋友站在这些材料前。问他们是否能看清你的材料?如果不能,那么,就要重新准备相关材料。

你的摘要必须简明扼要。剔除所有不必要的细枝末节。研究方法要用尽可能简明的语言来描述,并使用图来解释说明。根据你的研究,插图要能表现不同条件下的刺激材料,或是你调查中的代表性条目,或是你实验设计的流程。实验结果要用简单易读的符号和字母在图表中标示。如果可以,你应该将统计结果在图标题上标示出来。总之,你不应仅仅列出原始数据或统计结果的表格。最后,你要在结论部分简要概括你的实验结果,总结出不超过 5 条的结论。观众在看海报时不会做笔记,而且他们在一个会上要看很多海报,所以,你应该在海报上只突出少数重点,以

使观众会后能有印象。你还要注意将材料以简单易懂的方式呈现给观众。例如,当观众不仅有心理学家还有其他领域的学者时,你要很注意地避免心理学专业术语,而应该将你的结论与实际应用相联系。

出于审美上的原因,许多人用彩纸来为海报上的各个结构块镶边,但这不能做得太夸张。比起漂亮的图表,学者们更欣赏漂亮的数据。如果可以的话,展板要粘在一块硬板上。现在最常用的方式是将整个海报做成 1 到 2 张大的软纸片,它们可以卷成卷轴以方便携带。这样,当会议刚开始你还有点紧张忙乱时,你也无需为展板是否整齐且距离一致而担心。最后,预览一下你的海报!可能由于字体过大以至于不能正常阅读,也可能存在大量印刷错误。我见过一些享誉世界的研究者,使用十分专业的海报却存在错别字,以至于他不得不用圆珠笔修改这些错误。这些研究者当时十分尴尬,而换作你也会这样。

你也要向同事提前展示一下海报,要他们提问题。在回答了这些问题后,要求他们对海报和呈现方式提改进建议。最后,如果可能的话,你可以将论文的书面完整版分发给大家。他们会觉得这是个不错的做法。

无论你将如何展示自己的研究,都应该为之自豪并做一些有效的工作。要知道,这是你目前所有工作的结晶。如果你以一个不清楚或没意思的方式呈现自己的工作成果,那么,之前的所有努力都白费了。

小　结

一项研究,只有被其他的研究者所了解才有价值。研究者需要通过写一个高水平的实验报告来让他人明白自己的研究。这一报告需要遵循 *APA Publication Manual*。心理学报告撰写所遵循的原则又不同于历史或文学评论。例如,用词更直白,较少使用直接引用的语句,副标题用词更中性,注脚离题的情况更少,很少将知识争论转为人身攻击,往往不武断地下结论。一篇报告应包括标准化的环节。由于很多读者是靠读标题来决定是否阅读整篇报告,因此,标题要简短并包括足够多的信息,以帮助读者确定是否需要读你这篇报告。作者和单位名写在标题之后。摘要(约 120 字)是对完整报告的简短概括,应以它作为正文前书页的结束。

在报告的正文中,引言要回顾大量的文献,以使读者了解现今相关知识和本研究的目的。方法部分要提供尽量详细的信息,使读者能重复你的实验步骤。它一般分为 3 个部分:被试(介绍被试的特征、人数及招募途径);仪器与材料(提供必要信息使读者能准备和实验一样的设备和材料);步骤(详细介绍对每位被试进行的实验处理)。结果部分总结实验结果。这一部分包括描述性统计,以文字、表格和图的形式表现数据的集中和离散趋势。接下来是推断统计分析的结果,通常要包括效应值(effect size)。相关结果在讨论部分中会与引言中的知识进行联系和比较。报告还包括了一个以字母顺序排列的参考文献列表。

在保证基本格式的前提下,为了尽可能有效地传达信息,不需要再以第三人称来撰写论文。提倡使用生动的词,也允许偶尔使用第一人称。在引言和方法部分多使用过去时态,而在结果和讨论部分更提倡使用现在时态。为了使报告尽可能简明,要避免啰唆的用语,还应采用其他读者对报告的评价来使报告达到更高的水平。

避免语言偏向是很重要的，为使研究报告精确，你需要遵循以下 3 条建议：被试是什么人就实事求是地描述其是什么人，在涉及两种性别时避免用男性属性的词（如用 man 表示 human being，用 he 表示 he or she），用 participant 而不是 subject；询问人们他们想被怎样称呼，询问少数民族被试用怎样的术语称呼才合适；人是名词，而其属性是形容词，"People with disabilities"不是"the disabled"，"female participants"是"women"，而不是"females"。

研究还可以在专业会议上以会议论文和海报的形式报告。一篇论文一般由作者在 10～20 分钟内向 20～100 位同行呈现。它要以交流的形式呈现书面报告中的许多细节。为使观众能轻松地看到你的报告内容，你要精心准备如 PPT 这样的可视化内容。特别是步骤和结果部分，要用心揣摩听众目前所处的水平再作准备。报告者要对问题有所准备。海报要做成更为互动的形式。一个海报要包括标题和 6～9张海报纸，要对在房间内观看并提出问题的观众作一些讲解。海报要用比一般文本材料更大的字体，图要能看得清楚，但不要有过多的细节。会议前，论文和海报最好都在同事面前做好预讲。

结 语

我们探索不止,
我们探索的终点,
将是我们出发的地方,
而此时我们才真正知道这是哪里。

T. S. Eliot

祝贺你随着我做心理学实验的思路学习到这里。希望我的语言和图示不仅不会妨碍你的进步,还能引起你的兴趣。信息量与精确度之间的平衡很是微妙——这个平衡对于每个读者是不同的。我希望我的行文不会让你太烦。

很显然,这本书不会把你变成一个现成的实验心理学家,但我相信,它一定提供了足以使你做一些简单实验的信息。你将发现,做实验比阅读怎样做实验有趣得多。所以,现在来做一些有趣的事吧!

附录A

如何做基本的统计分析

如果你讨厌计算(见第 3 章),这个附录就是为你准备的。即使你只学过基础代数,你也能使用这个简单的"菜谱式"统计知识。这本书自然不是一本统计学书。一些使用本书的老师和学生,有时觉得需要对基本描述性统计和推断统计作一些简单介绍。现在我就告诉你一些检验方法。然而,我不准备告诉你为什么要这样做,我只能告诉你在什么条件下选择什么方法。

据我的观察,撰写与数字有关的内容,通常是很令人困扰的事情。因此,我试图通过一些例子告诉你如何处理数字。如果你按照例子中的方法和步骤整理数据,你就很少会遇到问题。

我首先介绍一下数据类型。然后,我会给你一个简短的统计符号词汇表。最后,我将向你介绍一些数据处理的实例。在每个例子的最后,我会提示你在论文中如何报告相应的结果。

数据类型

数字能够表达不同的信息。有时传达很多信息("到市场的距离是 28 公里"),有时则很少("第一位棒球手是 28 号")。有时描述和统计处理中会用到一些数字("这儿到剧院的距离是 14 公里,是到市场的一半")。而其他情况也使用同样的描述,就会很可笑("第二位棒球手是 14 号,因此他是第一位棒球手的一半")。因此,在你对数字作统计处理前,你一定要确定针对所用数字的类型,相应的统计处理是否合适。

称名量表

数字只是简单地用于命名,这就是称名量表。称名数据没有量化属性。你能做的合理的统计处理仅仅是计算相同属性的个体的数目,如有多少球员是 28 号?



图 A-1　各种数据类型的统计分析方法

顺序量表

数字能被排序或分等级，这就是顺序量表。前一年的赛车冠军能在他或她的汽车上标 1 号。第二名的可以标 2 号，以此类推。我们可以通过顺序数值获知 1 号车手的成绩好于 2 号车手，但是，我们无法知道差异大小。1 号车手和 2 号车手可能相差 500 分，2 号和 3 号可能仅仅相差 2 分。

等距量表(interval scale)

如果数字的间隔大小是有意义的，就称为等距量表。比如，华氏温度就是一个等距量表。50 °F 和 60 ℉之间的差，与 60 ℉和 70 ℉之间的差是相同的，都是 10 ℉。

等比量表

如果你能计算两个数的比例，并且这个比值是有意义的，可以比较的，就称为等比量表。虽然你不能说“20 ℉是 10 ℉的两倍”，但您能说“20 英里是 10 英里的两倍”。等距量表和等比量表最大的区别就是后者存在绝对零点(有实际意义的零点)。例如，对于华氏温度“零度”没有特别的意义，而将 32 °F 定为水的冰点。而如距离、身高、体重、体积都属于这类数据，就具有绝对零点。

在你作统计处理之前，首先需要回答一个问题“我处理的是什么类型的数据?”图 A-1 列出了在此附录中我们所提到的统计操作和相对应的数据类型。称名量表

是测量层次中较低的，而等比量表则是较高层次。因此，图中的阴影部分显示，对于等距和等比量表，可以进行所有的统计运算，而对于称名量表，仅能进行其中 3 种统计处理。

统计符号

X = 观察值
N = 个体数目
$\sum$ = 求和符号
$X^2 = X$ 的平方
$X^3 = X$ 的 3 次方
$\sqrt{X} = X$ 的平方根
$|X| = X$ 的绝对值

描述性统计

集中量数

众　数

众数是出现频次最多的数。计算一下每个值出现的次数，然后挑出出现次数最多的值。下面所举的例子中，众数是 2，因为它出现两次。

中　数

中数是中值，中点的数。首先，将所有数据排序，当数据个数为奇数时，中数就是排在中间的那个值。当数据个数为偶数时，中间两个数的均值为中数。下面的例子中，因为中间两个数是 2 和 3，所以中数是 2.5。

平均数

数据的总和除以数据的个数。Mean = M = $\overline{X} = \dfrac{\sum X}{N}$

例子：

X
1
2
2　$\overline{X} = \dfrac{17}{6} = 2.8$
3
4
5

$\sum \overline{X} = \overline{17}$

$N = 6$

论文中写成：$M = 2.8$

离散量数

全　距

全距是最大值和最小值之间的差值。在之前的例子中：

全距 $= 5 - 1 = 4$

方　差

$$方差 = S^2 = \frac{\sum (X - \overline{X})^2}{N}$$

例子：

	X	$\overline{X}^*$	$X - \overline{X}$	$(X - \overline{X})^2$
	1	3	-2	4
6 个值，	2	3	-1	1
因此，$N=6$	3	3	0	0
	3	3	0	0
	4	3	1	1
	5	3	2	4
	$\sum X = \overline{18}$			$\sum (X - \overline{X})^2 = \overline{10}$

$* \text{ Mean} = \overline{X} = \frac{18}{6} = 3$

$$\frac{\sum (X - \overline{X})^2}{N} = \frac{10}{6} = 1.67$$

标准差

$$标准差 = SD = \sigma = \sqrt{S^2} = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$$

在之前的例子中：

$SD = \sqrt{1.67} = 1.29$

论文中写成：SD = 1.29

相关关系

列联系数

列联系数(C)用来测量列联表(R * C 表)中两个称名变量的关联程度。首先,要做卡方(χ^2)检验(见 315 页)。假设两个表的卡方值已算得,你想知道这两个称名变量之间的关系强度。如果已知 $\chi^2=15$,$N=100$,那么,列联系数为:

$$C=\sqrt{\frac{\chi^2}{N+\chi^2}}=\sqrt{\frac{15}{100+15}}=\sqrt{.130}=.36$$

这个值无需再进行显著性检验,因为卡方检验中已经计算了显著性。

论文中写成:$C(N=100)=.36$

斯皮尔曼等级相关系数

斯皮尔曼等级相关系数(*rho*)用来测量两个顺序变量的关联程度。在这个例子中,每个被试都有两个得分(或秩次),两者的差异 d 已经算出。

例子:

被试	第一次测量的等级	第一次测量的等级	d	d^2	
Bill	4	4	0	0	
Jane	1	2	-1	1	
Bob	5	5	0	0	
Pete	2	3	-1	1	$N=5$
Mary	3	1	+2	4	
				$\sum d^2=\overline{6}$	

$$Rho=1-\frac{6\sum d^2}{N^3-N}=1-\frac{6(6)}{125-5}=1-\frac{36}{120}$$

$$=1-.3=.7$$

为了判断样本相关是来自于偶然,还是反映了真正的总体的相关关系,我们必须参考附录 B 中 *rho* 的临界值(见表 B-1)。我们可以看到,当 $N=5$ 时,*rho* 必须等于 1 才有显著性,例子的结果没有(*rho* = .7)。我们还可以从表中看出,假设存在相关关系,N 越大,越能检验出显著性。

论文中写成:$\rho(N=5)=.70, p>.05$

皮尔逊积差相关系数

皮尔逊积差相关系数(r)被用于测量两个等距或等比数据的相关强度。在下面的例子中,X 表示某个变量,Y 表示另一个变量。

例子:

被试	X	X^2	Y	Y^2	XY	
Tom	9	81	8	64	72	
Sue	4	16	4	16	16	
Jill	4	16	6	36	24	
Dave	2	4	4	16	8	$N=8$
Ken	1	1	3	9	3	
Jo	3	9	2	4	6	
Juan	7	49	8	64	56	
Al	5	25	5	25	25	
	$\sum X = 35$	$\sum X^2 = 201$	$\sum Y = 40$	$\sum Y^2 = 234$	$\sum XY = 210$	

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} = \frac{8(210) - (35)(40)}{\sqrt{8(210) - 35^2}\sqrt{8(234) - 40^2}}$$

$$= \frac{1\,680 - 1\,400}{\sqrt{1\,608 - 1\,225}\sqrt{1\,872 - 1\,600}} = \frac{280}{\sqrt{382}\sqrt{272}} = \frac{280}{(19.57)(16.49)}$$

$$= \frac{280}{322.7} = .868$$

为了检验这 8 对数据的 r 值是否有统计学意义，可以参考附录 B 中 r 的临界值（见表 B-2）。使用这个表格，你首先要确定自由度（df），自由度 $df=N-2$。因此，例子中 $df=6$，当 $\alpha=.01$ 时，表中的值是.834。因为.868 大于.834，所以 $p<.01$。也就是说，在 100 次中只有 1 次相关强度是由于抽样误差造成的。

论文中写成：$r(6)=.87, p<.01$

推断性统计

卡方检验

卡方检验用于判断期望频数与观察频数之间是否具有统计学差异。

例子：

	硬币多次投出反面后猜正面向上的被试数目	硬币多次投出反面后猜反面向上的被试数目
观察值(O)	60	40
期望值(E)	50	50
$O-E$	+10	−10
$(O-E)^2$	100	100
$\frac{(O-E)^2}{E}$	2	2

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 2 + 2 = 4$$

期望频数是根据经验频数分布或者某种理论频数分布计算出来的频数。通常，理论频数是观察频数的或然预期。比如上述例子中，预期值指的是被试预测没有任何偏差（或说没有赌徒谬误“gambler's fallacy”）——一半机会正面向上，一半机会反面向上。

作推断统计的最后一步是将你最后计算出的结果和临界值表进行比较。卡方的临界值见附录 B 中的表 B-3。查表之前，你要按照以下方法判断自由度：

df = 分组数目 -1　本例 $df = 2 - 1 = 1$

在界值表中我们找到自由度为 1，$\chi^2 > 3.84$ 时，$p < .05$。因此，本例的数据在 .05水平上具有统计学差异。另外，$\chi^2 = 4 < 6.64$，因此，在 .01 水平差异未达到显著。

论文中写成：$\chi^2(1,\ N = 100) = 4.00, p < 0.05$

独立样本 t 检验

t 检验有两种，一种用于独立测量数据，一种用于相关测量数据。独立样本 t 检验用于判定两个单独的样本之间的差异是否显著。假设以下数据为正态分布。

例子：

第 1 组			
X_1	$\bar{X}_1$	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$
9	7	2	4
8	7	1	1
7	7	0	0
7	7	0	0
4	7	-3	9
$\sum X_1 = \overline{35}$			$\sum (X_1 - \bar{X})^2 = \overline{14}$

$N_1 = 5$

$$M_1 = \bar{X}_1 = \frac{\sum X_1}{N_1} = \frac{35}{5} = 7$$

$$\sigma_1 = \sqrt{\frac{\sum (X_1 - \bar{X}_1)^2}{N_1}} = \sqrt{\frac{14}{5}} = \sqrt{2.8} = 1.67$$

第 2 组			
X_2	$\bar{X}_2$	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
5	3	2	4
4	3	1	1
3	3	0	0
2	3	-1	1
1	3	-2	4
$\sum X_2 = \overline{15}$			$\sum (X_2 - \bar{X}_2)^2 = \overline{10}$

$N_2 = 5$

$$M_2 = \overline{X}_2 = \frac{\sum X_2}{N_2} = \frac{15}{5} = 3$$

$$\sigma_2 = \sqrt{\frac{\sum (X_2 - \overline{X}_2)^2}{N_2}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.41$$

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{\sigma_1}{\sqrt{N_1 - 1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{N_2 - 1}}\right)^2}} = \frac{7 - 3}{\sqrt{\left(\frac{1.67}{\sqrt{5 - 1}}\right)^2 + \left(\frac{1.41}{\sqrt{5 - 1}}\right)^2}}$$

$$= \frac{4}{\sqrt{\left(\frac{1.67}{2}\right)^2 + \left(\frac{1.41}{2}\right)^2}} = \frac{4}{\sqrt{.697 + .497}} = \frac{4}{\sqrt{1.194}} = \frac{4}{1.09} = 3.67$$

独立样本的自由度计算如下：

$$df = N_1 + N_2 - 2$$
$$= 5 + 5 - 2 = 8$$

从附录 B 中表 B-4 可以看出，$df = 8$，$p < .01$，t 必须大于3.355。我们的结果是 $t = 3.67$，所以在.01 水平显著。

论文中写成：$t(8) = 3.67, p < .01$

相关样本 t 检验

相关样本 t 检验用于判断两个条件一致或匹配样本之间的差异（D）是否显著。

$$t = \frac{\overline{X}_D}{\frac{\sigma_D}{\sqrt{N - 1}}}$$

例子：

被试	条件1	条件2	差值(D)	$\overline{X}_D^*$	$X_D - \overline{X}_D$	$(X_D - \overline{X}_D)^2$
1	9	6	3	3	0	0
2	8	5	3	3	0	0
3	7	5	2	3	-1	1
4	8	4	4	3	1	1
5	8	5	3	3	0	0
$N = \overline{5}$	$\sum X_1 = \overline{40}$	$\sum X_2 = \overline{25}$	$\sum D = \overline{15}$			$\sum (X_D - \overline{X}_D)^2 = \overline{2}$

$$* M_D = \overline{X}_D = \frac{\sum D}{N} = \frac{15}{5} = 3$$

$$\sigma_D = \sqrt{\frac{\sum (D - \overline{D})^2}{N_2}} = \sqrt{\frac{2}{5}} = \sqrt{.4} = .632$$

$$t = \frac{\overline{X}_D}{\frac{\sigma_D}{\sqrt{N - 1}}} = \frac{3}{\frac{.632}{\sqrt{5 - 1}}} = \frac{3}{\frac{.632}{2}} = \frac{3}{.316} = 9.49$$

相关样本的自由度计算如下：

$$df = N - 1 = 5 - 1 = 4$$

从附录 B 中表 B-4 可以看出，本例中 t 必须大于 4.604，才能达到 $p<.01$ 水平。本例 $t=9.49$，所以在.01 水平显著。

论文中写成：$t(4)=9.49, p<.01$

Mann-Whitney U 检验

Mann-Whitney U 检验使用条件和独立样本 t 检验一致，但要求数据是非正态分布或非等距的。

$$\left.\begin{aligned} U &= N_1N_2+\frac{N_1(N_1+1)}{2}-R_1 \\ &\text{或} \\ U &= N_1N_2+\frac{N_2(N_2+1)}{2}-R_2 \end{aligned}\right\} U\text{ 值以较小者为准}$$

其中：

N_1：U 值较小组的样本量

N_2：U 值较大组的样本量

R_1：U 值较小组的秩次总和

R_2：U 值较大组的秩次总和

例子：

	第 1 组			第 2 组	
	X_1	秩次		X_2	秩次
	1	1		2	2
	3	3.5		4	5
	3	3.5		7	8
	5	6		8	9.5
$N_1=10$	6	7	$N_2=10$	10	13.5
	8	9.5		13	16
	9	11.5		15	17
	9	11.5		16	18
	10	13.5		17	19
	12	15		18	20
		$R_1=82$			$R_2=128$
					Note

$$U=N_1N_2+\frac{N_1(N_1+1)}{2}-R_1=(10)(10)+\frac{10(10+1)}{2}-82$$

$$=100+\frac{110}{2}-82=73$$

或

$$U=(10)(10)+\frac{10(10+1)}{2}-128=27$$

两个 U 值因为 27 较小，所以最后 $U=27$。

附录 B 中表 B-5，B-6 为 U 值临界值表。如果想看 $p<.05$ 水平的 U 值，我们可以看表 B-5，当 $N_1=10$，$N_2=10$ 时，$U=23$。为了达到显著性，我们所得的值必须等于或小于这个临界值。所求 $U=27$，不满足条件，所以在.05 水平未检出显著性。

注意，Mann-Whitney U 检验与其他检验不同，计算所得 U 值必须小于而不是大于临界值，才表示结果有显著性意义。当 $N_1<7$ 或者 $N_2>20$ 时，要选择更高级的临界值表，U 值必须按照以下公式转换成 z 值：

$$z=\frac{U-\frac{N_1N_2}{2}}{\sqrt{\frac{(N_1)(N_2)(N_1+N_2+1)}{12}}}$$

z 值的临界值见附录 B 的表 B-7。

论文中写成：$U(N_1=10,N_2=10)=27,p>.05$

Wilcoxon 符号秩和检验

Wilcoxon 符号秩和检验用于判断两个条件一致或匹配样本之间的差异（D）是否显著。它和相关样本 t 检验不同之处在于，它可以用于非正态分布数据或顺序数据。

$$\left.\begin{aligned}T&=\sum R_+\\T&=\left|\sum R_-\right|\end{aligned}\right\}\text{以较小者为准}$$

其中：

R_+ 是指正差异的秩次

R_- 是指负差异的秩次

例子：

配对数	条件1	条件2	D	不论正负的秩次	正差异秩次	负差异秩次
1	54	50	4	3	3	
2	47	32	15	9	9	
3	39	33	6	4	4	
4	42	45	-3	2.5		-2.5
5	51	38	13	7	7	
6	46	39	7	5	5	
7	42	44	-2	1		-1
8	54	46	8	6	6	
9	42	39	3	2.5	2.5	
10	47	33	14	8	8	
					$\sum R_+=\overline{45.5}$	$\left\|\sum R_-\right\|=\overline{3.5}$

3.5 小于 45.5，因此：

$$T=\left|\sum R_-\right|=3.5$$

临界值可查阅附录 B 的表 B-8。为了达到相应的显著性，T 值必须等于或小于临界值。本例有 10 对数据，因此 $n=10$，假设我们不知道两种条件差异的方向，我们使用双尾检验。可以看到，$3<3.5<5$，因此 $p<.02$。

论文中写成：$T(n=10)=3.50,p<.02$

方差分析

方差分析（ANOVA）用于近似正态分布的等距或等比数据。ANOVA 检验可用

于被试内设计(重复测量),也可用于被试间设计(独立测量)以及多变量设计。然而,在这个附录中,仅讨论单变量的被试间设计。下面的例子中,自变量有三个水平,所列的公式适合于多于3个组别的设计。

虽然ANOVA的计算看上去似乎很复杂,但其检验原理是相当简单的。假设你做一个实验,你要从3个组中收集数据。实验目的在于测量3个样本是否来自同一总体,其差异仅来源于偶然(取样误差),还是三者来自不同总体,其差异来自处理效应和偶然误差。在ANOVA中,你可以基于样本分布进行方差(变异)分解。变异可分解为两个基本部分,一个是处理间方差(包括处理组间的差异和个体差异),另一个是处理内方差(偶然样本误差引起的差异)。

在ANOVA中最后要计算的是F值,即组间变异和组内变异比较得出的一个比值。如果样本来自同一总体,当处理没有效应时,这个比值就很接近于1。也就是说,组间差异与组内差异大致相等。然而,当处理效应存在时,样本来自于不同总体,那么,组间差异就会大于组内差异。F值就大于1。F值越大,越能说明组间差异是来自于处理效应,而非偶然误差。

接下来的例子中,我们首先计算组间平方和(SS_{bg})、组内平方和(SS_{wg})以及总平方和(SS_{TOT})。然后我们用SS_{bg}和SS_{wg}分别除以相应的自由度,而得到组间均方(MS_{bg})和组间均方(MS_{wg})。最后用MS_{bg}除以MS_{wg},即得F值。

你应该能理解以下的例子,如果有困难,先理解以下几个定义可能会有帮助:

T:所有组所有数据的总和

T_j:每组(组j)数据的总和,下标j表示组序($j=1,2,3$)

N:所有组数据的总个数

n_j:每组中(组j)数据的个数

$\sum_{j=1}^{k}$:从1到k组的数据之和

k:组数,即处理条件的个数

例子:

第1组		第2组		第3组	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
3	9	9	81	10	100
5	25	6	36	8	64
4	16	5	25	11	121
3	9	8	64	10	100
1	1	7	49	9	81
2	4	7	49	10	100
5	25	6	36	11	121
2	4	4	16	12	144
3	9	8	64	10	100
1	1	7	49	9	81
$T_1=29$	103	$T_2=67$	467	$T_3=100$	1 012
$n_1=10$		$n_2=10$		$n_3=10$	

$N=10+10+10=30$

$T = 29 + 67 + 100 = 196$

$k = 3$

$$SS_{TOT} = \sum X^2 - \frac{T^2}{N} = (103 + 469 + 1\,012) - \frac{(196)^2}{30}$$

$$= 1\,584 - \frac{38\,416}{30} = 1\,584 - 1\,281 = 303$$

$$SS_{bg} = \sum_{j=1}^{k} \frac{T_j^2}{n_j} - \frac{T^2}{N} = \frac{29^2}{10} + \frac{67^2}{10} + \frac{100^2}{10} - \frac{(196)^2}{30}$$

$$= \frac{841}{10} + \frac{4\,489}{10} + \frac{10\,000}{10} - 1\,281$$

$$= 84.1 + 448.9 + 1\,000 - 1\,281 = 1\,533 - 1\,281 = 252$$

$$SS_{wg} = SS_{TOT} - SS_{bg} = 303 - 252 = 51$$

$df_{bg} = k - 1 = 3 - 1 = 2$

$df_{wg} = N - k = 30 - 3 = 27$

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}} = \frac{252}{2} = 126$$

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}} = \frac{51}{27} = 1.89$$

$$F = \frac{MS_{bg}}{MS_{wg}} = \frac{126}{1.89} = 66.7$$

现在我们可以在附录 B 的表 B-9 查看 F 值的临界值。分子的 $df = 2$，分母的 $df = 27$，若 $p < .05$，F 要大于等于 3.38，若 $p < .01$，F 要大于等于 5.57。本例中 $F = 66.7$，远大于这些临界值，所以，组间有显著性差异。注意，只要任意两组间有显著性差异，检验结果即有显著性差异。如果想知道两两之间的具体差异状况，要作进一步的检验——这个内容超出了本书的范围，可供参考的书目在第 12 章的结尾提到了一些。

论文中写成：$F(2,27) = 66.70, p < .05$

Kruskal-Wallis 检验

如果是非等距、等比数据或非正态分布数据，那么，就要使用 Kruskal-Wallis 检验方法来判断两组或多组间的差异。数据必须是顺序变量。

在下面例子中：

K：组数，即处理条件的个数

n_j：每组中数据的个数，下标 j 表示组序（$j = 1, 2, 3$）

N：所有组数据的总个数

R_j：每组（组 j）秩次的总和

t：某一相同秩次所含数据的个数

例子：

第 1 组		第 2 组		第 3 组	
X_1	等级	X_2	等级	X_3	等级
8	15	2	2.5	6	11
4	5.5	5	8.5	5	8.5
7	13	2	2.5	4	5.5
5	8.5	3	4	5	8.5
7	13	1	1	7	13
	$R_1 = 55.0$		$R_2 = 18.5$		$R_2 = 46.5$

$K = 3$

$n_j = 5$

$N = 15$

给所有组别中的每个得分编秩次。

得分	秩次	平均秩次	t
1	1	1	
2	2	2.5	2
2	3		
3	4	4	
4	5	5.5	2
4	6		
5	7	8.5	4
5	8		
5	9		
5	10		
6	11	11	
7	12	13	3
7	13		
7	14		
8	15	15	

现在将编好的秩次填入之前的那个表格中，然后，计算出 R_1, R_2, R_3。

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1)$$

$$= \frac{12}{15(15+1)}\left[\frac{(55)^2}{5} + \frac{(18.5)^2}{5} + \frac{(46.5)^2}{5}\right] - 3(15+1)$$

$$= \frac{12}{15(16)}\left[\frac{3\,025}{5} + \frac{342.25}{5} + \frac{2\,162.25}{5}\right] - 3(16)$$

$$= \frac{12}{240}\left[\frac{5\,529.5}{5} - 48\right]$$

$$= .05(1\,105.9) - 48 = 55.295 - 48 = 7.295$$

校正 H 值，用 H 值除以 $1 - \frac{\sum(t^3 - t)}{N^3 - N}$.

$$1 - \frac{(2^3-2) + (2^3-2) + (4^3-4) + (3^3-3)}{15^3 - 15}$$

$$1-\frac{(8-2)+(8-2)+(64-4)+(27-3)}{3\,375-15}$$

$$1-\frac{96}{3\,360}=1-.029=.971$$

$$H=\frac{7.295}{.971}=7.51$$

根据附录 B 中表 B-10 可以查阅临界值。在上述例子中，每组容量分别是 5，5 和 5，$H=7.51$ 时，$p<.049$. 因此，本例中组间差异在.05 水平有显著性。由于 7.51 <7.98，所以，在.01 水平上差异无显著性。

如果组容量超过 5，那么，H 值的分布近似于卡方分布。本例临界值的判断也可以参考表 B-3，相应自由度为 $k-1$。

论文中写成：$H(5,5,5)=7.51, p<.05$

结 论

本附录列出了一些基本的统计操作。然而，如果你想做一些更复杂的实验，你需要至少再做三件事情。第一，你需要学习使用一些复杂的检验方法——多变量设计或被试间被试内混合设计；第二，你需要学习一些统计软件以节约时间并提高效率；第三，也是最重要的一点，你需要学习的不仅仅是一点点统计基础知识，作为一个研究者，你需要理解你为什么要这样做。

理解统计操作中的各种概念，不仅有助于你选择最有效力的方法分析你的数据，而且还可以帮助你进行研究设计，以便你能有效利用数据。统计顾问最害怕这样的情景——无经验的研究者拿着一列数据来到他的办公桌前问到："我如何分析这些数据？"在某些情况下，数据无法分析。

问题的关键在于实验设计和统计是相互联系的。如果你准备设计研究，你就应该理解将要使用的统计方法的基本概念。

附录B
统计表

表 B-1　斯皮尔曼等级相关(*rho*)的临界值

N	$p = .050\,0$	$p = .010\,0$
5	1.000	—
6	.886	1.000
7	.786	.929
8	.738	.881
9	.683	.833
10	.648	.794
12	.591	.777
14	.544	.715
16	.506	.665
18	.475	.625
20	.450	.591
22	.428	.562
24	.409	.537
26	.392	.515
28	.377	.496
30	.364	.478

来源:Olds E G. 小样本(N < 30)等级差离均平方和的分布 *Annals of Mathematical Statistics*, 1938, 9, 133-148;等级差离均平方和与校正的离均平方和的5%显著性水平 *Annals of Mathematical Statistics*, 1949, 20, 177-118。表 B-1 取自 Elementary Statistics, Underwood et al. , Appleton-Century-Crofts。

表 B-2 皮尔逊积差相关(r)的临界值

df	双尾检验的显著性水平		
	.10	.05	.01
1	.988	.997	.999 9
2	.900	.950	.990
3	.805	.878	.959
4	.729	.811	.917
5	.669	.754	.874
6	.622	.707	.834
7	.582	.666	.798
8	.549	.632	.765
9	.521	.602	.735
10	.497	.576	.708
11	.476	.553	.684
12	.458	.532	.661
13	.441	.514	.641
14	.426	.497	.623
15	.412	.482	.606
16	.400	.468	.590
17	.389	.456	.575
18	.378	.444	.561
19	.369	.433	.549
20	.360	.423	.537
25	.323	.381	.487
30	.296	.349	.449
35	.275	.325	.418
40	.257	.304	.393
45	.243	.288	.372
50	.231	.273	.354
60	.211	.250	.325
70	.195	.232	.303
80	.183	.217	.283
90	.173	.205	.267
100	.164	.195	.254

来源:取自 Fisher R A. *Statistical Methods for Research Workers*, 14th Edition. 1973, Habner Press.

表 B-3 χ^2 分布表

df	$p = .05$	$p = .01$
1	3.84	6.64
2	5.99	9.21
3	7.82	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.80
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.88	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

来源：Fisher & Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Longman Group Ltd., London(先前由 Oliver and Boyd Ltd., Edinburgh 出版)。获得版权许可。

表 B-4 t 分布表(双尾概率)

df	$p=.10$	$p=.05$	$p=.02$	$p=.01$
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
60	1.671	2.000	2.390	2.660
∞	1.645	1.960	2.326	2.576

来源:选自 Fisher & Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London(原来由 Oliver and Boyd Ltd., Edinburgh 出版),获得作者和出版社的版权许可。

表 B-5 $p < .05$ 时的 Mann-Whitney U 值的临界值

N_2	N_1													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

来源：改编自 *Bulletin of the institute of Educational Research at Indiana University*, 1953,1(2)中的表 1,3,5 和 7,以及 Mann-Whitney 扩展表。

表 B-6 $p < .01$ 时的 Mann-Whitney U 值的临界值

N_2	N_1													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	—	—	0	0	0	1	1	1	2	2	2	2	3	3
4	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	24	30	36	42	48	54	60	67	73	79	86	92	99	105

来源：改编自 *Bulletin of the institute of Educational Research at Indiana University*, 1953,1(2)中的表 1,3,5 和 7,以及 Mann-Whitney 扩展表。

表 B-7 Z 分数表

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.500 0	.496 0	.492 0	.488 0	.484 0	.480 1	.476 1	.472 1	.468 1	.464 1
.1	.460 2	.456 2	.452 2	.448 3	.444 3	.440 4	.436 4	.432 5	.428 6	.424 7
.2	.420 7	.416 8	.412 9	.409 0	.405 2	.401 3	.397 4	.393 6	.389 7	.385 9
.3	.382 1	.378 3	.374 5	.370 7	.366 9	.363 2	.359 4	.355 7	.352 0	.348 3
.4	.344 6	.340 9	.337 2	.333 6	.330 0	.326 4	.322 8	.319 2	.315 6	.312 1
.5	.308 5	.305 0	.301 5	.298 1	.294 6	.291 2	.287 7	.284 3	.281 0	.277 6
.6	.274 3	.270 9	.267 6	.264 3	.261 1	.257 8	.254 6	.251 4	.248 3	.245 1
.7	.242 0	.238 9	.235 8	.232 7	.229 6	.226 6	.223 6	.220 6	.217 7	.214 8
.8	.211 9	.209 0	.206 1	.203 3	.200 5	.197 7	.194 9	.192 2	.189 4	.186 7
.9	.184 1	.181 4	.178 8	.176 2	.173 6	.171 1	.168 5	.166 0	.163 5	.161 1
1.0	.158 7	.156 2	.153 9	.151 5	.149 2	.146 9	.144 6	.142 3	.140 1	.137 9
1.1	.135 7	.133 5	.131 4	.129 2	.127 1	.125 1	.123 0	.121 0	.119 0	.117 0
1.2	.115 1	.113 1	.111 2	.109 3	.107 5	.105 6	.103 8	.102 0	.100 3	.098 5
1.3	.096 8	.095 1	.093 4	.091 8	.090 1	.088 5	.086 9	.085 3	.083 8	.082 3
1.4	.080 8	.079 3	.077 8	.076 4	.074 9	.073 5	.072 1	.070 8	.069 4	.068 1
1.5	.066 8	.065 5	.064 3	.063 0	.061 8	.060 6	.059 4	.058 2	.057 1	.055 9
1.6	.054 8	.053 7	.052 6	.051 6	.050 5	.049 5	.048 5	.047 5	.046 5	.045 5
1.7	.044 6	.043 6	.042 7	.041 8	.040 9	.040 1	.039 2	.038 4	.037 5	.036 7
1.8	.035 9	.035 1	.034 4	.033 6	.032 9	.032 2	.031 4	.030 7	.030 1	.029 4
1.9	.028 7	.028 1	.027 4	.026 8	.026 2	.025 6	.025 0	.024 4	.023 9	.023 3
2.0	.022 8	.022 2	.021 7	.021 2	.020 7	.020 2	.019 7	.019 2	.018 8	.018 3
2.1	.017 9	.017 4	.017 0	.016 6	.016 2	.015 8	.015 4	.015 0	.014 6	.014 3
2.2	.013 9	.013 6	.013 2	.012 9	.012 5	.012 2	.011 9	.011 6	.011 3	.011 0
2.3	.010 7	.010 4	.010 2	.009 9	.009 6	.009 4	.009 1	.008 9	.008 7	.008 4
2.4	.008 2	.008 0	.007 8	.007 5	.007 3	.007 1	.006 9	.006 8	.006 6	.006 4
2.5	.006 2	.006 0	.005 9	.005 7	.005 5	.005 4	.005 2	.005 1	.004 9	.004 8
2.6	.004 7	.004 5	.004 4	.004 3	.004 1	.004 0	.003 9	.003 8	.003 7	.003 6
2.7	.003 5	.003 4	.003 3	.003 2	.003 1	.003 0	.002 9	.002 8	.002 7	.002 6
2.8	.002 6	.002 5	.002 4	.002 3	.002 3	.002 2	.002 1	.002 1	.002 0	.001 9
2.9	.001 9	.001 8	.001 8	.001 7	.001 6	.001 6	.001 5	.001 5	.001 4	.001 4
3.0	.001 3	.001 3	.001 3	.001 2	.001 2	.001 1	.001 1	.001 1	.001 0	.001 0
3.1	.001 0	.000 9	.000 9	.000 9	.000 8	.000 8	.000 8	.000 8	.000 7	.000 7
3.2	.000 7									
3.3	.000 5									
3.4	.000 3									
3.5	.000 23									
3.6	.000 16									
3.7	.000 11									
3.8	.000 07									
3.9	.000 05									
4.0	.000 03									

注：表中是 H_0 条件下 z 值对应的单尾概率。最左边一列是不同的 z 值。最上面一行是不同的值。因此，$z \geq .11$ 或 $z \leq -.11$ 对应的 $p = .456\ 2$。

来源：经 Siegel S. *Nonparametric Statistics for the Behavioral Science.* New York McGraw-Hill Book Company，1956（p247）版权许可。

表 B-8 Wilcoxon 符号秩和检验的 T 临界值(两样本比较的秩和检验用)

	单尾检验的显著性水平					单尾检验的显著性水平			
	.05	.025	.01	.005		.05	.025	.01	.005
	双尾检验的显著性水平					双尾检验的显著性水平			
n	.10	.05	.02	.01	n	.10	.05	.02	.01
5	0	—	—	—	28	130	116	101	91
6	2	0	—	—	29	140	126	110	100
7	3	2	0	—	30	151	137	120	109
8	5	3	1	0	31	163	147	130	118
9	8	5	3	1	32	175	159	140	128
10	10	8	5	3	33	187	170	151	138
11	13	10	7	5	34	200	182	162	148
12	17	13	9	7	35	213	195	173	159
13	21	17	12	9	36	227	208	185	171
14	25	21	15	12	37	241	221	198	182
15	30	25	19	15	38	256	235	211	194
16	35	29	23	19	39	271	249	224	207
17	41	34	27	23	40	286	264	238	220
18	47	40	32	27	41	302	279	252	233
19	53	46	37	32	42	319	294	266	247
20	60	52	43	37	43	336	310	281	261
21	67	58	49	42	44	353	327	296	276
22	75	65	55	48	45	371	343	312	291
23	83	73	62	54	46	389	361	328	307
24	91	81	69	61	47	407	378	345	322
25	100	89	76	68	48	426	396	362	339
26	110	98	84	75	49	446	415	379	355
27	119	107	92	83	50	466	434	397	373

注:T 是所有符号相同差异的等级之和的较小值。对一个 n 值而言,如果它小于等于表中的值,那么,T 就是显著的。

来源:Roger E K, *Elementary Statistics*, 2nd Edition. Pacific Grove, CA: Brooks/Cole, 1984.

表 B-9 F 分布表

分母自由度	分子自由度									
	1	2	3	4	5	6	8	12	24	∞
1	161.45	199.50	215.72	224.57	230.17	233.97	238.89	243.91	249.04	254.32
	4 032.10	4 999.03	5 403.49	5 625.14	5 764.08	5 859.39	5 981.34	6 105.83	6 234.16	6 366.48
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	9.33	6.93	5.93	5.41	5.06	4.82	4.50	4.16	3.78	3.36
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.01	2.57
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01

表 B-9 *F* 分布表(续表)

分母自由度	分子自由度									
	1	2	3	4	5	6	8	12	24	∞
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.52
	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.82
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44
	7.17	5.06	4.20	3.72	3.41	3.19	2.89	2.56	2.18	1.68
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.89	1.67	1.35
	7.01	4.92	4.07	3.60	3.29	3.07	2.78	2.45	2.07	1.53
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.88	1.65	1.31
	6.98	4.88	4.04	3.56	3.26	3.04	2.74	2.42	2.03	1.47
90	3.95	3.10	2.71	2.47	2.32	2.20	2.04	1.86	1.64	1.28
	6.92	4.85	4.01	3.53	3.23	3.01	2.72	2.39	2.00	1.43
100	3.94	3.09	2.70	2.46	2.30	2.19	2.03	1.85	1.63	1.26
	6.90	4.82	3.98	3.51	3.21	2.99	2.69	2.37	1.98	1.39
200	3.89	3.04	2.65	2.42	2.26	2.14	1.98	1.80	1.57	1.14
	6.97	4.71	3.88	3.41	3.11	2.89	2.60	2.28	1.88	1.21
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	
	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	

注:上一行数字是.05 水平上检验的值,下一行数字是.01 水平上检验的值。

来源:改编自 Garret H E. *Statistics in Psychology and Education*, 5th Edtion. Copyright 1958, David McKay Co. Inc 中的表 *F*。

表 B-10 Kruskal-Wallis H 值的临界值(三样本比较的秩和检验用)

样本大小					样本大小				
n_1	n_2	n_3	H	p	n_1	n_2	n_3	H	p
2	1	1	2.700 0	.500	4	3	2	6.444 4	.008
2	2	1	3.600 0	.200				6.300 0	.011
2	2	2	4.571 4	.067				5.444 4	.046
			3.714 3	.200				5.400 0	.051
3	1	1	3.200 0	.300				4.511 1	.098
3	2	1	4.285 7	.100	4	3	3	4.444 4	.102
			3.857 1	.133				6.745 5	.010
3	2	2	5.357 2	.029				6.709 1	.013
			4.714 3	.048				5.790 9	.046
			4.500 0	.067				5.727 3	.050
			4.464 3	.105				4.709 1	.092
3	3	1	5.142 9	.043	4	4	1	4.700 0	.101
			4.571 4	.100				6.666 7	.010
			4.000 0	.129				6.166 7	.022
3	3	2	6.250 0	.011				4.966 7	.048
			5.361 1	.032				4.866 7	.054
			5.138 9	.061				4.166 7	.082
			4.555 6	.100	4	4	2	4.066 7	.102
			4.250 0	.121				7.036 4	.006
3	3	3	7.200 0	.004				6.872 7	.011
			6.488 9	.011				5.454 5	.046
			5.688 9	.029				5.236 4	.052
			5.600 0	.050				4.554 5	.098
			5.066 7	.086	4	4	3	4.445 5	.103
			4.622 2	.100				7.143 9	.010
4	1	1	3.571 4	.200				7.136 4	.011
4	2	1	4.821 4	.057				5.598 5	.049
			4.500 0	.076				5.575 8	.051
			4.017 9	.114				4.545 5	.099
4	2	2	6.000 0	.014	4	4	4	4.477 3	.102
			5.333 3	.033				7.653 8	.008
			5.125 0	.052				7.538 5	.011
			4.458 3	.100				5.692 3	.049
			4.166 7	.105				5.653 8	.054
4	3	1	5.833 3	.021	5	1	1	4.653 9	.097
			5.208 3	.050	5	2	1	4.500 1	.104
			5.000 0	.057				3.857 1	.143
			4.055 6	.093				5.250 0	.036
			3.888 9	.129				5.000 0	.048
								4.450 0	.071
								4.200 0	.095
								4.050 0	.119

表 B-10 Kruskal-Wallis H 值的临界值(续表)

样本大小					样本大小				
n_1	n_2	n_3	H	p	n_1	n_2	n_3	H	p
5	2	2	6.533 3	.008	5	4	4	7.760 4	.009
			6.133 3	.013				7.744 0	.011
			5.160 0	.034				5.657 1	.049
			5.040 0	.056				5.617 6	.050
			4.373 3	.090				4.618 7	.100
			4.293 3	.122				4.552 7	.102
5	3	1	6.400 0	.012	5	5	1	7.309 1	.009
			4.960 0	.048				6.836 4	.011
			4.871 1	.052				5.127 3	.046
			4.017 8	.095				4.909 1	.053
			3.840 0	.123				4.109 1	.086
5	3	2	6.909 1	.009				4.036 4	.105
			6.821 8	.010	5	5	2	7.338 5	.010
			5.250 9	.049				7.269 2	.010
			5.105 5	.052				5.338 5	.047
			4.650 9	.091				5.246 2	.051
			4.494 5	.101				4.623 1	.097
5	3	3	7.078 8	.009				4.507 7	.100
			6.981 8	.011	5	5	3	7.578 0	.010
			5.648 5	.049				7.542 9	.010
			5.515 2	.051				5.705 5	.046
			4.533 3	.097				5.626 4	.051
			4.412 1	.109				4.545 1	.100
5	4	1	6.954 5	.008				4.536 3	.102
			6.840 0	.011	5	5	4	7.822 9	.010
			4.985 5	.044				7.791 4	.010
			4.860 0	.056				5.665 7	.049
			3.987 3	.098				5.642 9	.050
			3.960 0	.102				4.522 9	.099
5	4	2	7.204 5	.009				4.520 0	.101
			7.118 2	.010	5	5	5	8.000 0	.009
			5.272 7	.049				7.980 0	.010
			5.268 2	.050				5.780 0	.049
			4.540 9	.098				5.660 0	.051
			4.518 2	.101				4.560 0	.100
5	4	3	7.444 9	.010				4.500 0	.102
			7.394 9	.011					
			5.656 4	.049					
			5.630 8	.050					
			4.548 7	.099					
			4.523 1	.103					

来源:经出版社和作者同意,改编自 Kruskal W H and Wallis W A. Use of ranks in one-criterion variance analysis, *Journal of American Statistical Association*, 1952, 47,614-617。(该表的校正来自作者的勘误表,*Journal of American Statistical Association*, 1953, 48,910.)

附录C

随机数字表

行\列	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	43342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253

续表

行＼列	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65797	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83317	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
46	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
47	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
48	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
49	95012	68379	93526	70765	10593	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653

来源：105000 随机数表，声明号 4914，文件号 261-A-1，Interstate Commerce Commission，Washington，DC，May 1949.

术语

ABBA平衡(ABBA counterbalancing):这是用来消除只有A、B两个水平的单因素实验中存在变量线性混淆效应的技术。先呈现水平A,再呈现两个水平B,最后,再呈现水平A。

横坐标(Abscissa):图的水平坐标轴,即x轴。它通常用于代表自变量的不同水平。

摘要(Abstract):研究报告的一个简短总结,字符最多在960个以内。

变式信度(Alternative-form reliability):通过一个与第一个测验有相似项目的测验对同一组人进行测量,用两次测验分数之间的相关系数确定测验信度的方法。

模拟理论(Analogical theory):通过模拟物理模型的方法解释心理关系的理论。

方差分析(Analysis of variance):用于分析析因实验或多水平单因素实验结果的参数统计方法。

应用研究(Applied research):以寻找解决特定问题方法为主要目的的研究。

档案法(Archival research):一种对现存的公开或隐私文件的检验、组织和解释的研究方法。

渐近线(Asymptote):一个减加速方程逐渐接近坐标轴时画出的一条虚拟线。

条形图(Bar graph):一种用间隔的垂直条棒表示数据频次的方法。横坐标表示类别,条棒的高度表示不同类别的频次。

基线实验(Baseline experiment):一种单变量实验,其效应可以从单个被试数据中表现出来。首先,建立一个稳定反应基线;随后,施加实验处理并形成转换状态;最后,撤销实验处理,基线再次恢复。

基础研究(Basic research):以探讨科学基本原理为主要目的的研究。虽然有些研究也可以解决一些实际应用问题,但其目的还是为了丰富人们的知识体系。

组间设计(Between-subjects design):一种实验研究的策略,在这种实验设计中,每个被试只接受一个水平的处理条件。

双峰分布(Bimodal distribution):频率分布中有两个隆起,每个隆起都有一个最多频次。

盲实验(Blind experiment):实验中被试并不知道自己所接受的是哪一个处理条件。

序列效应(Carry-over effect):组内设计中的一种效应,即先前的处理水平对后来的反应的影响。

个案法(Case history):一种非实验研究方法,是对个体或单个事件的行为数据的分析。

天花板效应(Ceiling effect):对一个分布顶端数据的剪切,这是由于最高分数的限制造成的。

卡方检验(Chi-square test):一种非参数检验的统计推论方法。它用于检验一种事件出现的频率与另一个现象出现频率之间的差异性。

选择反应时(Choice reaction time):从众多刺激中选择一个反应的时间。

封闭式问题(Close-ended question):一种调查问题,它要求调查对象按照设定好的结构回答。

决定系数(Coefficient of determination):相关系数的平方,它是被解释方差的百分比。

完全平衡法(Complete counterbalancing):一种实验设计方法。在这种设计中,自变量的每个水平在其他水平前后的次数相同。

复合因变量(Composite dependent variable):将许多因变量综合在一个因变量之中,它能够反应被试的整体反应。

同时效度(Concurrent validity):这是一种测验效度,它能够检验测验与效标同时实施时,测验对效标的预测程度。

会议论文(Conference paper):学术会议上向与会者口头报告的研究计划。

混淆变量(Confounding variable):一个变量水平与自变量水平之间相互关联的变量,这使得研究者无法辨别实验效应是自变量引起,还是其他变量引起。

内容效度(Content validity):一种测验效度,通过仔细分析测验中的内容,然后,构建测验,从而获得更具代表性的测验项目。

上下文效应(Context effect):自变量的不同水平对被试行为的影响。其中一个是实验中的自变量,另一些则是实验中出现的其他变量。

可能性系数(Contingency coefficient):衡量两个称名变量之间联系强度的指标。

控制组(Control group):在组间设计中,一组被试不接受任何实验处理,它被作为与接受实验条件被试的相比较的组。

控制变量(Control variable):实验者为了防止实验效应的变异而设置的特定实验水平。

聚合序列设计(Converging-series design):一系列按顺序进行实验,用于排除其他可能的假设。

相关(Correlation):两个变量之间具有特定的方向和强度关系。

相关研究(Correlational observation):研究者为了确定两个或多个变量之间的关系而设计的一种研究,它不会对变量进行控制,也不能进行因果推论。

相关系数(Correlation coefficient):用于表示两个变量之间关系的方向和强度的数值,它在 -1 和 1 之间。

·**平衡**(Counterbalancing):一种用于减少或消除混淆变量顺序效应的方法。它是对不同实验处理条件呈现的顺序进行调整。

关键事件(Critical incident):在应用研究中,被作为诊断自变量和因变量之间关系的单个事件。

交叉相关(Crossover interaction):析因设计中的交互作用,在图中表示自变量的两条线相交叉。

曲线方程(Curvilinear function):一种非直线的方程,它包含了很多可以用各种不同数学公式表达的成分。

欺骗(Deception):向被试隐瞒实验目的。

演绎(Deduction):一种从一组前提中得出结论的方法。得出的结论不会超过前提能够提供的信息,如 A 是 B,B 是 C;那么,A 是 C。

需求特性(Demand characteristics):一个实验应该具有的特性,它可以使被试按一定方式进行反应,通常是为了支持实验假设,与自变量水平无关。

因变量(Dependent variable):实验者选择测量的行为,这些行为可能依赖于自变量水平的变化。

描述统计(Descriptive statistics):减少数据量的一种方法,它可以只描述数据的主要特征。例如,集中趋势和离中趋势等。

描述性理论(Descriptive theory):只是简单将名称列上,而不去解释事件发生的原因或过程。

方向性问题(Directional hypothesis):在相关研究中,不知道哪个变量是原因,哪个变量是结果。

指向性问题(Directionality problem):无法知道哪一个变量是原因,哪些效应来自相关关系。

双盲实验(Double-blind experiment):在实验中,被试和主试都不知道所呈现的是自变量的哪一个水平。

双任务方法(Dual-task methodology):一种从同时完成的第二个任务中推测实验效应的方法。

电子出版(Electronic publishing):用电子的方式而不是以纸质的方式呈现信息的方法,如互联网。

民族志(Ethnography):描述一种文化的定性研究方法。

实验组(Experimental group):在组间设计中接受处理条件的组。

实验方法(Experimental method):一种操纵自变量,同时测量因变量的研究方法。实验方法可以进行因果推论:因变量的任何变化都是有自变量操纵引起的。

外部效度(External validity):实验结果在现实世界中的总体、条件或环境的可推广程度。

结果造假(Fabrication of results):一种通过虚假数据伪造科学结果的方式。

表面效度(Face validity):一种最弱的测验效度,它只是简单地从表面判断测验是否能够测量到所要测量的东西。

析因设计(Factorial design):在一个实验中自变量不只一个,并且不同自变量的各个水平间相互组合。

因素(Factors):在析因设计中包含不同水平的自变量。

地板效应(Floor effect):对一个分布低端数据的剪切,这是由于最低分数的限制造成的。

频率分布(Frequency distribution):不同类别事件发生的频次在一个图中画出来。

方程(Function):描述两个变量之间关系的一条直线或曲线。

线性实验(Functional experiment):实验的自变量有 3 个或更多水平,以便确定自变量和因变量之间的线性关系。

集体施测(Group administration of surveys):一组人同时完成调查的收集数据方法。

霍桑效应(Hawthorne effect):行为的变化是由研究者对被试的关注引起,而不是由其他因素引起。

直方图(Histogram):一种用相连的竖条描述定量数据频次的方法。横坐标表示数据类别,竖条的高度表示每一个类别的频次。

影响内部效度的历史因素(History as a threat to internal validity):因变量的变化归咎于自变量两个水平测验之间发生的事件。

假设(Hypothesis):对两个或多个变量之间关系的预测性描述。

自变量(Independent variable):研究者所操纵的 2 个或多个条件,以确定因变量的效应。

演绎(Induction):一种逻辑推理过程,在此过程中结论所包含的信息比用于得出结论的材料多。

推论统计(Inferential statistic):一种将观测结果归结于某种变化的统计检验方法。

知情同意(Informed consent):一种确保参与研究的人充分了解研究的重要信息并同意参加实验的过程。

研究评估委员会(Institutional review board):研究机构中的一个委员会,它的功能是保证研究都符合伦理。

析因设计中的交互作用(Interaction in a factorial design):多个自变量不同组合在因变量上的不可加效应。

选择的交互作用(Interactions with selection):由成熟、历史因素与选择因素之间产生的交互作用,它会影响内部效度。

内部效度(Internal validity):衡量因变量的变化是由操纵自变量引起的确定程度。

网络调查(Internet survey):通过互联网进行的调查。

间断序列设计(Interrupted time-series design):一种准实验设计,是对一个组在实验操作前后分别进行若干次观测。

等距量表(Interval scale):数据间距相等的测量指标,$1 = n - (n - 1)$。例如,华氏温度。

访谈(Interview):通过与被访者面对面交流收集结构化或非结构化数据的调查方法。

Krukal-Wallis 等级单向方差分析(Kruskal-Wallis one-way ANOVA by rank):对不同实验组的等级数据进行检验的推论统计方法。

拉丁方(Latin square):一种确保自变量不同水平在不同位置上出现次数相等的平衡方法。

显著性水平(Level of significance):从一个样本得到的结果是因为机遇引起而不是真实发生的概率,这个概率通常是 $p < 0.05$ 或 $p < 0.01$。

Likert 量表(Likert scale):通过受访者对某一话题的赞同程度而了解他们的态度。

线性方程(Linear function):一条直线方程。

线图(Line graph):用连续的直线或曲线描绘连个变量之间关系的方法。

文献搜索(Literature search):对与某个领域相关的以往研究结果进行搜索的过程。

通讯调查(Mail survey):通过信件开展的调查。

主效应(Main effect):在析因设计中,自变量不同水平对应的因变量之间的关系。

Mann-Whitney *U* 型检验(Mann-Whitney *U* test):一种非参数推论统计方法。它用于检验不同组的等级数据之间关系。

匹配组设计(Matched-groups design):一种组间设

计方法。它按照与因变量高度相关的变量将被试进行匹配,然后,将它们随机分到两组中。

影响内部效度的成熟因素(Maturation as a threat to internal validity):因变量的变化由被试在不同处理之间的年龄变化或经验增加而引起。

平均数(Mean):数据集中趋势的指标。通过将所有数据之和除以数据的个数计算。

平均处理效应(Mean treatment effect size):衡量实验操纵对行为影响效应大小的指标。计算方法是实验组的平均数减去控制组的平均数,再除以控制组的标准差。

中数(Median):一种集中趋势指标。计算方法是将所有数据排序,位于中间的那个数就是中数。

元分析(Meta-analysis):从众多实验中总结概括实验累计效应的方法。

混合析因设计(Mixed factorial design):一种至少有一个组内变量和一个组间变量的析因设计。

众数(Mode):一种集中趋势指标。出现频次最多的那个数就是众数。

单调函数(Monotonic function):一个变量总是随着另一个增加而增加或减少的函数。

影响内部效度的个人因素(Mortality as a threat to internal validity):因变量的变化是由于在不同自变量条件下被试的特点引起的。

多选问题(Multiple-alternative question):限定了选择范围的问题。

自然观察(Naturalistic observation):在自然情境中开展的研究。

负函数(Negative function):在一个函数中,一个变量增加,而另一个变量减少。

负加速函数(Negative accelerated function):在一个函数中,一个变量随着另一变量的增加而增加或减少。这样的函数在开始时陡峭,而后来逐渐变得平缓。

称名量表(Nominal scale):一种不具备定量特性的量表,它只是用数字代表名称。例如,一个别在跑步运动员运动服上的号码342。

无差别转换(见对称性转换)(Nondifferential transfer):在一个实验中自变量有A和B两个水平,A水平后面的B水平对行为的效应等同于B水平后面的A水平对行为的效应。

无指向性假设(Nondirectional hypothesis):一种试探性预测。自变量的不同水平将引起因变量的改变,但不会对这种改变的方向进行预测。

不等同控制组设计(Nonequivalent control group design):一种准实验设计,它所设定的控制组包含了与实验组不同的特点。

非实验设计(Nonexperimental design):一种不同于实验设计或准实验设计的实验设计,它不能克服影响内部效度的因素,因此,这种实验的结果比较脆弱。

非实验控制组(Nonexperiment control group):一个用于估计实验属性的被试。并不是每个被试都接受自变量的不同水平,但他们会被告知各种条件并回答他们将如何反应。

非单调函数(Nonmonotonic function):函数的斜率从正到负或从负到正至少变化一次。

非参数检验(Nonparimetric test):一种对未知总体分布情况下的推论统计方法,如是否是正态分布。

非反应性偏差(Nonresponse bias):不同组被试的反应速度差异造成的调查结果分布特点。

正态分布(Normal distribution):一种用特定数学方程得出的频次分布,它呈单峰对称的铃形,它的平均数、中数和众数相等。

虚无假设(Null hypothsis):虚无统计检验中必须的陈述,它指明总体中不存在由自变量不同水平引起的效应,因此,样本中的任何差异都是由机遇引起的。

单组后测设计(One-group posttest-only design):一种非实验设计,只有一个实验组接受自变量的一个水平。

单组前后测设计(One-group pretest-posttest design):一种非实验设计,先对一组被试进行测试,然后,接受一种实验条件,最后,再进行重测。

开放式问题(Open-ended question):一种可以让被调查者自由回答的问题。

操作定义(Operational definition):一种用来说明对概念进行操作和测量的具体方法。

顺序效应(Order effect):在组内设计中,测量到的行为变化依赖于自变量不同水平的呈现顺序。

顺序量表(Ordinal scale):一种测量量表。量表的数字顺序有意义,而数字间的间隔和比率之间没有任何意义,如9大于8。

纵坐标(Ordinate):图的纵坐标,它的大小通常代表因变量的不同水平。

页眉(Page header):每页右上角开头的几个字。

参数检验(Parametric test):一种推论统计检验。在这种检验中,对整体分布有一个假设,通常假设它是正态分布。

参数检验(Parametric test):一种推论统计方法,它假设总体成正态分布。

部分平衡法(partial counterbalancing):一种安排自变量不同水平呈现顺序的方法,以减少顺序混淆变量的干扰效应。

参与者(Participants):行为被研究者研究的人,也叫被试。

Pearson 极差相关系数(Pearson product-moment correlation coefficient):衡量两个等距或等比变量之间关系强度的统计量。

节省百分比(Percent savings):一种复合因变量。它是初学次数减去重学次数,再除以初学次数,并乘以100;它表示再一次学习能够节省的百分比。

预备实验(Pilot experiment):一个不能满足所有实验条件的小规模实验,其目的是对正式实验的水平和过程进行前测。

安慰剂(Placebo):在药物研究中,一种非药物像药物一样使用,这种非药物就是安慰剂。有时安慰剂可以像药物一样起作用,尽管它不是真正的药物。

剽窃(Plagiarism):不恰当地引用他人的话或观点。

正函数(Positive function):一个变量随着另一个变量的增加而增加的函数关系。

正加速函数(Positively accelerated function):一个变量的增加或减少的速度随另一个变量的增加而增大。函数的基本特性是开始阶段斜率较小,随着一个变量的增加,斜率越来越大。

张贴(Poster presentation):通过一系列面板展示研究计划的方法。

非等组后测设计(Posttest-only design with non-equivalent groups):一种非实验设计。其中一组接受自变量的一个水平,另一组接受自变量的第二个水平。

统计功效(Statistical power):虚无假设为伪时接受它的概率。

预测效度(Predictive validity):建立测验效度的一种方法。它通过测验是否能够预测某种效标来确定。

代理前测(Proxy pretest):一种与后测结果相关的测验,它被应用在含有不对等组的准实验设计中,其目的在于证实组与组之间的等质性。

近似问题(Proximate questions):科学中关于行为是如何发生的而不是为什么发生的问题。

心理历史(Psychohistory):通常是关于一些名人的心理传记。通过他们生活中的一些关键事件,了解并解释他们的行为模式。

PsycARTICLES:美国心理学会刊物中文章的网络服务数据库。

PsyINFO:美国心理学会刊物中论文摘要的网络搜索服务。

纯研究(见基础研究)(Pure research):用于了解科学基本规律的研究。尽管其研究结果也可以解决一些实际应用问题,但它的目的只是为了增强科学知识体系。

定性化设计(Qualitative designs):用描述性数据进行的研究设计,数据包括对人的描述,如对事件和环境的意见、态度等。

定量化设计(Quantitative designs):事件可以量化的研究设计,它的数据是数字,如实验数据。

定量化理论(Quantitative theory):用数学语言描述的关系的理论。

准实验设计(Quasi-experiment designs):实验中被试无法随机选择,但很多威胁内部效度的因素可以被排除。

问卷(Questionnaire):可以以单人或团体方式进行的书面调查。

随机分配(Random assignment):用随机的方式从总体中选择被试,并将他们随机分配到不同组中。

随机化(Randomization):一种选择的方法,它可以使每个项目被选的机会相同。

区组随机化(Randomization within blocks):按照每个区组中某种试验出现次数相同的原则,将试验分配到不同条件中的方法。

有限条件随机化(Randomization within constraints):按照某个或某些选择规则选择项目的方法。例如,按照每个项目出现次数相等的规则随机选择不同条件。

随机样本(Random sample):按照随机过程从总体中选择的样本。

随机选择(Random selection):采用随机过程从总体中选择项目或人的方法。

随机变量(Random variable):实验中水平由机遇而不是实验者控制的变量。

全距(Range):一组数中的最大值与最小值之间的差值。

全距效应(Range effect):在组内设计中,刺激或反应的顺序恒定,由于学习迁移的作用,最好成绩出现在中间位置。

评价量表(Rating scale):被试用等级顺序方式表示自己的评价。例如,从"最同意"到"最反对"。

比率量表(Ratio scale):数字的比值具有实际意义的量表。例如,4 厘米是 2 厘米的 2 倍。

信度(Reliability):测量能够被重复的程度。

重复测量设计(见组内设计)(Repeated measure design):被试接受所有自变量水平的实验设计。

反应-表面方法(Response-surface methodology):用于估计多个自变量组合效应的技术,这种方法不需实施完整的析因实验。

翻转性(Reversibility):在基线实验中,实验操作撤销后,原来的状态出现恢复的现象。

散点图(Scatterplot):一种用图描述数据的方法。图中的每个点对应于横坐标上变量的值。

二手文献(Secondary source):研究报告中引用的未阅读的参考文献,但这些参考文献来源的第一手文献则是阅读过的。

影响内部效度的选择因素(Selection as a threat to internal validity):由于不同被试接受不同的实验条件而造成的因变量变化。

独立组设计(见组间设计)(Separate groups design):每一个被试组只接受一种自变量的一个水平。

模拟控制组(Simulation control group):要求一组被试模拟做出接受实验条件下的行为,以便估计实验的特性。

偏态分布(Skewed distribution):一个不对称的分布,它一边的尾部比另一边长。

小 *N* 基线设计(见基线实验)(Small-*N* baseline design):一种单变量实验,它可以从少量被试的数据中获得实验效应。首先,建立一个反应基线;然后,实施实验处理,并建立转换状态;最后,撤除实验处理,基线状态恢复。

Spearman 等级相关系数(Spearman rank-order correlation coefficient):衡量两个等级变量之间关系强度的方法。

分半信度(Split-half reliability):确定测验信度的一种方法。它是将测验结果分成两半,并计算它们之间的相关。

标准差(Standard deviation):一个分布的离中趋势统计量。它的计算方法是,将每个分数与平均数相减,然后,再计算这些差值的平方和;平方和再除以分数的数量,最后,再求平方根。

统计结论效度(Statistical conclusion validity):衡量自变量和因变量之间存在显著关系的程度。

影响内部效度的统计回归(Statistical regression as a threat to internal validity):在重复测验中,极端分数向平均数移动的现象。

统计显著性(Statistical significance):当判断样本的效应不可能是机遇引起,而是由样本所在总体的特性决定的概率符合科学家们设定的概率时,那么,就可以说结果达到了统计显著性。这个概率通常是 $p<0.05$ 或 $p<0.01$。

稳定状态(Steady state):在基线实验的开始阶段,反应量的变化很微弱。

分层取样(Stratified sampling):按照某种特性将总体分组,然后,从不同组中抽取相应的被试。例如,收入水平、族群等。

调查(Survey):通过询问人们的意见和行为的研究方法。

对称迁移(Symmetrical transfer):在有 A 和 B 两个水平的自变量的实验中,B 在 A 后面的效应与 A 在 B 后面的效应是相同的。

电话调查(Telephone survey):通过电话访谈的调查方法。

影响内部效度的测验过程(Testing as a threat to internal validity):因变量的变化是被试先前参加的测验工具或情景引起。

重测信度(Test-retest reliability):确定测验信度的方法。它是通过对同一组实施两次相同的测试,然后,计算两次测验分数之间的相关系数。

理论(Theory):对一组抽象变量之间关系的陈述。

第三变量问题(Third variable problem):在相关研究中无法判断第一个变量是否引起第二个变量的变化,也不能判断是否有第三个变量引起了其他两个变量的变化。

三向交互作用(Three-way interaction):一种高阶交互作用。当每个双向交互作用的性质都依赖于第三个因素的不同水平时,就出现了三向交互作用。

稳态转换(Transition steady state):反应速度不会因实验条件的变化而改变。

处理(Treatment):实验者使用的实验操作,通常与控制条件进行比较。

处理×被试设计(Treatment × subject design):见被试内设计。

剪裁的分布(Truncated distribution):对有限频率分布的变量范围进行限制。例如,天花板或地板效应。

***t* 检验**(*t*-test):一种参数推论统计。它是用来确定观察到的自变量两个水平之间的差异由机遇引起的概率。

双向交互作用(Two-way interaction):在析因实验中,一个变量不同水平之间的差异依赖于另一个变量的不同水平。

Ⅰ类错误(Type Ⅰ error):当虚无假设为真时,拒绝它的概率。

Ⅱ类错误(Type Ⅱ error):当虚无假设为假时,未能拒绝它的概率。

终结问题(Ultimate questions):在科学研究中,有关行为发生原因的问题。

效度(Validity):某种事物(测量工具或概念)与标准一致的程度。

方差(Variance):一个分布的离中趋势指标,它是由每个分数与平均数的差值的平方和,再除以分数的数目。

Wilcoxon 配对符号等级检验(Wilcoxon matched-pairs signed-ranks test):一种适合于等级数据的推论统计方法。它是用来确定同一或匹配的被试的差异由机遇引起的概率。

组内设计(Within-subject design):每一个被试接受所有的实验处理条件或自变量水平的设计。

***x*-轴**(*x*-axis):见横坐标。图的横坐标,通常用来代表自变量的不同水平。

***y*-轴**(*y*-axis):见纵坐标。图的纵坐标,通常用来表示因变量的不同水平。

参考文献

AAU statement on preventing and probing research fraud. (1983). *Chronicle of Higher Education*, 26, 8.

Adair, J. G. (1973). *The human subject.* Boston: Little, Brown.

Adair, J. G., & Epstein, J. (1968). Verbal cues in the mediation of experimenter bias. *Psychological Reports*, 22, 1045-1053.

Adams, J. A. (1972). Research and the future of engineering psychology. *American Psychologist*, 27, 615-622.

Aldridge, J. W. (1978). Levels of processing in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 164-177.

American Medical Association. (1989, April). *AMA surveys of physician and public opinion on health care issues*: 1989. Chicago: Author.

American Psychological Association. (1997). Animal rights activity increases: Threats made against behavioral scientists. *Science Agenda*, 10, 1, 4.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, D. C.: Author.

American Psychological Association. (2002). *Mastering APA style: Instructor's resource guide.* Washington, D. C.: Author.

American Psychological Association. (2002). *Mastering APA style: Student's workbook and training guide.* Washington, D. C.: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060-1073.

American Psychological Association. (2005). *Concise rules of APA style.* Washington, D. C.: Author.

American Psychological Association. (2007). Guidelines for ethical conduct in the care and use of animals. Retrieved January 17, 2007, from http://www. apa. org / science /anguide. html.

Arnoult, M. D. (1972). *Fundamentals of scientific method in psychology.* Dubuque, IA: William C. Brown.

Associated Press. (1985). [Data available from POLL computer database]. Storrs, CT: Roper Center for Public Opinion.

Barber, B. (1976). The ethics of experimentation with human subjects. *Scientific American*, 234, 25-31.

Barber, T. X., & Silver, J. J. (1968). Fact, fiction, and the experimenter bias effect. *Psychological Bulletin. Monograph Supplement*, 70, 1-29.

Barber, T. X. (1976). *Pitfalls in human research.* New-York: Pergamon Press.

Best, J. (2001). *Damned lies and statistics: Untangling numbers from the media, politicians, and activists.* Berkley, CA: University of California Press.

Blampied, N. M. (2000). Single-case research designs: A neglected alternative. *American Psychologist*, 55, 960.

Boyce, J. R. (1989). Use of animals in research: Can we find a middle ground? *Journal of the Veterinary Medicine Association*, 194, 24-25.

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice-Hall.

Bryant, F. B., & Wortmen, P. M. (1978). Secondary analysis: The case for data archives. *American Psychologist*, 33, 381-387.

Burd, S. (1993). Report finds animal rights activists have stepped up attacks. *The Chronicle of Higher Education*, 40, A31.

Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking.* Englewood Cliffs, NJ: Prentice-Hall.

Carlyle, T. (1888). *On heroes, heroworship and the heroic in history.* New York: Fredrick A. Stokes & Brother. ii. Volume 5-Number 3. Editorial.

Carroll, M. E., & Overmier, J. B. (Eds.). (2001). *Animal research and human health: Advancing human welfare through behavioral science.* Washington, D. C.: American Psychological Association.

Christensen, L. (1988). Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin*, 14, 664-675.

Clark, C., & Williges, R. C. (1973). Response surface methodology central-composite design modifications for human performance research. *Human Factors*, 15, 295-310.

Coke-Pepsi slugfest. (1976, July 26). *Time*, pp. 64-65.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Cordes, C. (1990). U. S. enters lawsuit accusing scientist, institutions of fraud. *The Chronicle of Higher Education*, 37, A1, A24, A25.

Coren, S., & Halpern, D. F. (1991). Lefthandedness: A marker for decreased survival fitness. *Psychological Bulletin*, 109, 90-106.

Coren, S., & Porac,, C. (1977). Fifty centuries of right-

handedness: The historical record. *Science*, 198, 631-632.

Daly, M., & Wilson, M. (1988). *Homicide.* New York: Aldine de Gruyter.

Decker, B. (1967). Words about words: I. Pessimistic. *Journal of Creative Behavior*, 1, 34.

Dewsbury, D. A. (1990). Early interactions between animal psychologists and animal activists and the founding of the APA committee on precautions in animal experimentation. *American Psychologist*, 45, 315-327.

Diener, E., Lucas, R. E., & Scollon, C. N. (2006). Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist*, 61, 305-314.

Dillman, D. A. (1978). *Mail and telephone surveys: Tile total design method.* New York: Wiley.

Doyle, A. C. (1989). A scandal in Bohemia. In *The adventures of Sherlock Holmes.* New York: Tom Dougherty Associates. (Original work published 1891.)

Dunlap, K. (1920). *Mysticism, Freudianism, and Scientific Psychology.* St. Louis, MO: C. V. Mosby.

Erickson, F. (1973). What makes school ethnography "ethnographic"? *Anthropology & Education Quarterly*, 4, 10-19.

Estes, W. K. (1993). How to present visual information. APS *Observer*, 6(2), 6-9.

Feeney, D. M. (1987). Human rights and animal welfare. *American Psychologist*, 42, 593-599.

Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist*, 48, 1141-1147.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science*, 1, 115.

Foertsch, J., & Gernsbacher, M. A. (1997). In search of gender neutrality: Is singular *they* a cognitively efficient substitute for generic *he*? *Psychological Research*, 8, 106-111.

Gallup, G. G., & Suarez, S. D. (1985). Alternatives to the use of animals in psychological research. *American Psychologist*, 40, 1104-1111.

Gantt, W. H. (1928). Ivan P. Pavlov: A biographical sketch. In W. H. Gantt (Ed.), *I. P. Pavlov's lectures on conditioned reflexes* (pp, 11-31). New York: Liveright.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationalism and the concept of perception. *Psychological Review*, 63, 149-159.

Garvey, W. D., & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist*, 26, 349-362.

Gelfand, H., Walker, C. J., & American Psychological Association (2002). *Mastering APA style: Student's workbook and training guide.* Washington, D. C.: American Psychological Association.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Newbury Park, CA: Sage.

Glinski, R. J., Glinski, B. C., & Slatin, P. T. (1970). Nonnaivety contamination in conformity experiments: Sources, effects, and implications for control. *Journal of Personality and Social Psychology*, 16, 478-485.

Gosling, C., Knight, N., & McKenney, L. S. (Eds.). (1989). *Search PsycINFO: Student Workbook.* Washington, D. C.: American Psychological Association.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Greenberg, M. S. (1967). Role playing: An alternative to deception? *Journal of Personality and Social Psychology*, 7, 152-157.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314-320.

Grobe, R. P., Pettibone, T. J., & Martin, D. W. (1973). Effectiveness of lecture pace on noise level in a university classroom. *Journal of Educational Research*, 67, 73-75.

Gross, A. E., & Flemming, I. (1982). Twenty years of deception in social psychology. *Personality and Social Psychology Bulletin*, 8, 402-408.

Hadaway, C. K., & Marler, P. L. (1998). Did you really go to church this week? Behind the poll data. *The Christian Century*, 115, 472-476.

Halpern, D. F., & Coren, S. (1993). Left-handedness and life span: A reply to Harris. *Psychological Bulletin*, 114, 235-241.

Harcum, E. R. (1989). The highly inappropriate calibrations of statistical significance. *American Psychologist*, 44, 964.

Harlow, H. F. (1958). The nature of love. *American Psychologist*, 13, 673-685.

Harris, L. J. (1993a). Do left-handers die sooner than right-handers? Commentary on Coren and Hal-pern's (1991) "Left-handedness: A marker for decreased survival fitness." *Psychological Bulletin*, 114, 203-234.

Harris, L. J. (1993b). Reply to Halpern and Coren. *Psychological Bulletin*, 114, 242-247.

Havemann, J. (1989, July 13). Proposals on animal research rattle cages. *Albuquerque Journal*, p. E4.

Hayakawa, S. I. (1978, June-July). *Change*, 6.

Herzog, H. A. (1995). Has public interest in animal rights peaked? *American Psychologist*, 50, 945-947.

Hostetler, A. J. (1988, June). Indictment: Congress sends message on fraud. *APA Monitor*, p. 5.

Huff, D. (1954). *How to lie with statistics.* New York: Norton.

Human Factors and Ergonomics Society. (1995). *Instructions and guidelines for poster presenters: Guidelines for preparing and arranging posters.* Santa Monica, CA: Author.

Infeld, L. (1950). *Albert Einstein.* New York: Scribner's.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research*, 50, 438-460.

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1995). A history of facilitated communication: Science, pseudoscience, and antiscience. *American Psychologist*, 50, 760.

Jaschik, S. (1991). Agriculture Dept. issues final rules on care of research animals. *The Chronicle of Higher Education*, 37, A25-A31.

Jensen, A. R. (1978). Sir Cyril Burt in perspective. *American Psychologist*, 33, 499-503.

Johnson, D. A. (1971). Pupillary responses during a short-term memory task: Cognitive processing, arousal, or both? *Journal of Experimental Psychology*, 90, 311-318.

Jung, J. (1969). Current practices and problems in use of college students for psychological research. *Canadian Psychologist*, 10, 280-290.

Justice Department and Department of Agriculture release report on animal rights extremism activities. (1993, September). *Federation News*, p. 6.

Kanekar, S. (1990). Statistical significance as a continuum. *American Psychologist*, 45, 296.

Kennedy, J. E., & Landesman, J. (1963). Series effects in motor performance studies. *Journal of Applied Psychology*, 47, 202-205.

Kenrick, D. T., & Keefe, R. C. (1992). Age preferences in mates reflect sex differences in reproductive strategies. *Behavioral and Brain Sciences*, 15, 75-133.

Kihlstrom, J. F. (1995, June). From the subject's point of view: The experiment as conversation and collaboration between investigator and subject. Paper presented at the American Psychological Society Annual Convention.

Kimmel, A. J. (1996). *Ethical issues in behavioral research: A survey.* Cambridge, MA: Blackwell.

Kirk, R. E. (1968). *Experimental designs: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (1990). *Statistics: An introduction.* Fort Worth, TX: Holt, Rinehart & Winston.

Krasner, L. (1958). Studies of the conditioning of verbal behavior. *Psychological Bulletin*, 55, 148-170.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

LeCompte, M. D., & Preissle, J. (1993). *Ethnography and qualitative design in educational research.* San Diego, CA: Academic Press.

Lerner, I. M. (1968). *Heredity, evolution, and society.* San Francisco: W. H. Freeman.

Levenson, R. L., Jr. (1990). Comment on Harcum. *American Psychologist*, 45, 295-296.

Ley, W. (1955). *Salamanders and other wonders.* New York: Viking Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209.

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, 21, 1-3.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161-171.

Lorenz, K. Z (1962). *The editors of King Solomon's Ring.* Chicago: Time-Life Books.

Madigan, R. M., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist*, 50, 428-436.

Madson, L., & Hessling, R. M. (2001). Reader's perceptions of four alternatives to masculine generic pronouns. *The Journal of Social Psychology*, 141, 156-158.

Maggio, R. (1991). *The bias-free word finder: A dictonary of nondiscriminatory language.* Boston: Beacon Press.

Mangan, K. S. (1990). Universities beet up security at laboratories to protect researchers threatened by animal-rights activists. *The Chronicle of Higher Education*, 37, A16-A19.

Mann, C. (1990). Meta-analysis in the breech. *Science*, 249, 476-480.

Martin, D. W., & Kelly, R. T. (1974). Secondary task performance during directed forgetting. *Journal of Experimental Psychology*, 103, 1074-1079.

Masling, J. (1966). Role-related behavior of the subject and psychologist and its effects upon psychological data. In D. Levine (Ed.), *Nebraska symposium on motivation.* Lincoln: University of Nebraska Press.

McAskie, M. (1978). Carelessness or fraud in Sir Cyril Burt's kinship data? A critique of Jensen's analysis. *American Psychologist*, 33, 496-498.

McDonald, K. (1983). Fraud in scientific research: Is it the work of "psychopaths"? *Chronicle of Higher Education*, 26, 7.

Medvedev, Z. A. (1969). *The rise and fall of T. D. Lysenko* (I. M. Lerner, Trans.). New York: Columbia University Press.

Meyers, C. (1990). Association fights restrictions on use of animals in research, drawing praise from scientists and anger from its critics. *The Chronicle of Higher Educa-*

tion, 37, A25-A27.

Meyers, R. H. (1971). *Response surface methodology.* Boston: Allyn & Bacon.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371-378.

Miller, N. E. (1985). The value of behavioral research on animals. *American Psychologist*, 40, 423-440.

Mitchell, D. B., & Richman, C. L. (1980). Confirmed reservations: Mental travel. *Journal of Experimental Psyctology. Human Perception and Performance*, 6, 58-66.

Monte, C. F. (1975). *Psychology's scientific endeavor.* New York: Praeger.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design. *American Psychologist*, 56, 119-127.

Mueller, J. H., & Furedy, J. J. (2001). Reviewing for risk: What's the evidence that it works? APS Observer, 14 (7), 1.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review*, 86, 214-255.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *Amrican Psychologist*, 41, 1299-1301.

Nicks, S. D., Korn, J. H., & Mainieri, T. (1997). The rise and fall of deception in social psychology and personality research, 1921 to 1994. *Ethics and Behavior*, 7, 69-77.

Nicol, A. A. M., & Pexman, P. M. (1999). *Presenting your findings: A practical guide for creating tables.* Washington, D. C.: American Psychological Association.

Olson, P. O. (1989). Motorcycle conspicuity revisited. *Human Factors*, 31, 141-146.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776-783.

Orne, M. T. (1970). Hypnosis, motivation and the ecological validity of the psychological experiment. In W. J. Arnold & M. M. Page (Eds.), *Nebraska symposium on motivation.* Lincoln: University of Nebraska Press.

Orne, M. T., & Evans, F. J. (1965). Social control in the psychological experiment: Antisocial behavior and hypnosis. *Journal of Personality & Social Psychology*, 1, 189-200.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist*, 52, 746-747.

Palladino, J. J., & Handelsman, M. M. (1995). On the light side: The history of APA format and style. *Psy Chi Newsletter*, 21, 6.

Parsons, H. M. (1974). What happened at Hawthorne? *Science*, 183, 922-932.

Paxton, J. P. (1997). "Someone with like a life wrote it": The effects of a visible author on high school history students. *Journal of Educational Psychology*, 89, 235-250.

Pifer, L., Shimizu, K., & Pifer, R. (1994). Public attitudes toward animal research: Some international comparisons. *Society & Animals*, 2, 95-113.

Pirsig, R. M. (1975). *Zen and the art of motorcycle maintenance.* New York: Bantam.

Plous, S. (1996a). Attitudes toward the use of animals in psychological research and education: Results from a national survey of psychologists. *American Psychologist*, 51, 1167-1180.

Plous, S. (1996b). Attitudes toward the use of animals in psychological research and education: Results from a national survey of psychology majors. *Psychological Science*, 7, 352-358.

Popper, K. R. (1968). *The logic of scientific discovery* (Rev. ed.). New York: Basic Books.

Porac, C., & Coren, S. (1981). *Lateral preferences and human behavior.* New York: Springer-Verlag.

Position statement on the use of animals in research. (1993). *NIH Guide*, 22, 2-3.

Poulton, E. C. (1973). Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin*, 80, 113-121.

Poulton, E. C. (1979). Composite model for human performance in continuous noise. *Psychological Review*, 86, 361-375.

Poulton, E. C., & Freeman, P. R. (1966). Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin*, 66, 1-8.

Pressley, M. M., & Tullar, W. L. (1977). A factor interactive investigation of mall survey response rates from a commercial sample. *Journal of Marketing Research*, 14, 108-111.

Protecting Human Research Subjects: Institutional Review Board Guide (1993). *NIH Guide Grants Contracts*, 22.

Puglisi, T. (2001). IRB review: It helps to know the regulatory framework. APS *Observer*, 14 (5), 1.

Pyrczak, F. & Bruce, R. R. (Eds.). (2000). *Writing empirical research reports: A basic guide for students of the social and behavioral sciences.* Los Angeles: Pyrczak Publishing.

Reed, J. G., & Baxter, P. M. (2003). *Library use: A handbook for psychology* (3rd ed.). Washington, D. C.: American Psychological Association.

Roediger, R. (2004). What should they be called? APS *Observer*, 17 (4), 5, 46-48.

Roethlisberger, F. J. (1977). *The elusive phenomena: An autobiographical account of my work in the field of organized behavior at the Harvard Business School.* Cam-

bridge, MA: Division of Research, Graduate School of Business Administration (distributed by Harvard University Press).

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance test to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553-565.

Roscoe, S. M. (1980). *Aviation psychology*. Ames: Iowa State University Press.

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In T. Rosenthal & T. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press.

Rosenthal, R., & Fode, K. L. (1973). The effect of experimental bias on the performance of the albino rat. *Behavioral Science*, 8, 183-189.

Rosenzweig, S. E. G. (1970). Boring and the Zeitgeist: Eruditone gesta beavit. *Journal of Psychology*, 75, 59-71.

Rosnow, R. L., & Rosenthal, R. (1999). *Beginning behavioral research: A conceptual primer* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Rowland, L. W. (1939). Will hypnotized persons try to harm themselves or others? *Journal of Abnormal and Social Psychology*, 34, 114-117.

Ryder, R. D. (1979). The struggle against speciesism. In D. Paterson & R. D. Ryder (Eds.), *Animal rights—a symposium* (p. 14). London: Centaur Press.

Satire, W. (1979, November 4). On language; the fumblerules of grammar. *New York Times Magazine*, pp. 16-18.

Sales, B. D., & Folkman, S. (2000). *Ethics in research with human participants*. Washington, D. C.: American Psychological Association.

Schindler, G. E. (1967). Why engineers and scientists write as they do: Twelve characteristics of their prose. *IEEE Transactions on Engineering Writing and Speech*, *EWS*-10, 32.

Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin*, 72, 214-228.

Segal, E. (1982). Editorial. *Journal of the Experimental Analysis of Behavior*, 38, 115.

Seligman, M. E. P. (1988). President's comments. *American Psychological Association Monitor*, 29, 2.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sheppard, L. (2001). *Examining the effect of humor in text-based instruction on learner's enjoyment and recall*. Unpublished practicum paper.

Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.

Singer, P. (1976). *Animal liberation*. London: Jonathan Cape.

Singer, P. (1985). *In defense of animals*. New York: Basil Blackwell.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, 57, 193-216.

Skinner, B. F. (1953). *Science and human behavior*. New York: Free Press.

Skinner, B. f. (1959). A case history in scientific method. In S. Koch (Ed.), *Psychology: A study of a science*. New York: McGraw-Hill.

Skinner, B. F. (1966). Operant behavior. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application* (p. 21). New York: Appleton-Century-Crofts.

Smart, R. (1966). Subject selection bias in psychological research. *Canadian Psychologist*, 7, 115-121.

Staff (1983). *Chronicle of Higher Education*, 29, 7.

Staff. (1994, June 27). Vox pop. *Time*, p. 26.

Strunk, W., Jr., & White, E. B. (1979). *The elements of style* (3rd ed.). New York: Macmillan.

Tanur, J. M. (1994). The trustworthiness of survey research. *The Chronicle of Higher Education*, 40, B1-B3.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-N experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum.

Toor, R. (2006). What's to "enjoy"? *The Chronicle of Higher Education*, 52, B5.

Van Orden, G. C., & Paap, K. R. (1997). Functional neural images fail to discover pieces of mind in parts of the brain. *Philosophy of Science Journal*, 64, 885-994.

Wainer, H. (2000). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Mahwah, NJ: Lawrence Erlbaum.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 77, 273-295.

Westfall, R. S. (1973). Newton and the fudge factor. *Science*, 179, 751-758.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention*. New York: Academic Press.

Wilson, G. T. (1985). Limitations of meta-analysis in the evaluation of the effects of psychological therapy. *Clinical Psychology Review*, 5, 35-47.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, 17, 657-658.

Woodworth, R. S. (1940). *Psychology* (4th ed.). New York: Holt.

万卷方法总书目

万卷方法是我国第一套系统介绍社会科学研究方法的大型丛书,来自中国社科院、北京大学等研究机构和高校的两百余名学者参与了丛书的写作和翻译工作。至今已出版图书60多个品种,其中绝大多数是2007年以来出版的新书。

1 应用STATA做统计分析(更新至STATA10.0)
978-7-5624-4483-1

2 社会调查设计与数据分析——从立题到发表
978-7-5624-6074-9

3 质性研究导引
978-7-5624-6132-6

4 APA格式——国际社会科学学术写作规范手册
978-7-5624-6105-0

5 如何做心理学实验
978-7-5624-6151-7

6 话语分析导论
978-7-5624-6075-6

7 心理学学位论文写作全程指导
978-7-5624-6113-5

8 心理学研究方法导论
978-7-5624-5828-9

9 分类数据分析
978-7-5624-6133-3

10 结构方程模型:AMOS的操作与应用(附光盘版)
978-7-5624-5720-6

11 AMOS与研究方法(第2版)
978-7-5624-5569-1

12 爱上统计学(第2版)
978-7-5624-5891-3

13 社会科学定量研究的变量类型、方法选择与范例解析
978-7-5624-5714-5

14 案例研究:设计与方法(中译第2版)
978-7-5624-5732-9

15 问卷设计手册:市场研究、民意调查、社会调查、健康调查指南
978-7-5624-5597-4

16 广义潜变量模型:多层次、纵贯性以及结构方程模型
978-7-5624-5393-2

17 调查问卷的设计与评估
978-7-5624-5153-2

18 心理学论文写作——基于APA格式的指导
978-7-5624-5354-3

19 心理学质性资料的分析
978-7-5624-5363-5

20 问卷统计分析实务:SPSS操作与应用
978-7-5624-5088-7

21 如何做综述性研究
978-7-5624-5375-8

22 质性访谈方法
978-7-5624-5307-9

23 量表编制:理论与应用(校订新译本)
978-7-5624-5285-0

24 质性研究:反思与评论(第2卷)
978-7-5624-5143-3

25 实验设计原理:社会科学理论验证的一种路径
978-7-5624-5187-7

26 混合方法论:定性研究与定量研究的结合
978-7-5624-5110-5

27 社会统计学
978-7-5624-5253-9

28 校长办公室的那个人(质性研究个案阅读)
978-7-5624-4880-8

29 泰利的街角(质性研究个案阅读)
978-7-5624-4937-9

30 客厅即工厂(质性研究个案阅读)
978-7-5624-4886-0

31 标准化调查访问
978-7-5624-5062-7

32 解释互动论
978-7-5624-4936-2

33 如何撰写研究计划书
978-7-5624-5087-0

34 质性研究的理论视角:一种反身性的方法论
978-7-5624-4889-1

35 社会评估:过程、方法与技术
978-7-5624-4975-1

36 如何解读统计图表
978-7-5624-4906-5

37 公共管理定量分析:方法与技术(第2版)
978-7-5624-3640-9

38 量化研究与统计方法
978-7-5624-4821-1

39 心理学研究要义
978-7-5624-5098-6

40 调查研究方法(校订新译本)
978-7-5624-3289-0

41 分析社会情境:质性观察和分析方法
978-7-5624-4690-3

42 建构扎根理论:质性研究实践指南
978-7-5624-4747-4

43 参与观察法
978-7-5624-4616-3

44 文化研究:民族志方法与生活文化
978-7-5624-4698-9

45 质性研究方法:健康及相关专业研究指南
978-7-5624-4720-7

46 如何做质性研究
978-7-5624-4697-2

47 质性研究中的访谈:教育及社会科学研究者指南
978-7-5624-4679-8

48 案例研究方法的应用(中译第2版)
978-7-5624-3278-3

49 教育研究方法论探索
978-7-5624-4649-1

50 实用抽样方法
978-7-5624-4487-9

51 质性研究:反思与评论(第1卷)
978-7-5624-4462-6

52 社会科学研究的思维要素(第8版)
978-7-5624-4465-7

53 哲学史方法论十四讲
978-7-5624-4446-6

54 社会研究方法
978-7-5624-4456-5

55 质性资料的分析:方法与实践(第2版)
978-7-5624-4426-8

56 实用数据再分析法(第2版)
978-7-5624-4296-7

57 质性研究的伦理
978-7-5624-4304-9

58 叙事研究:阅读、倾听与理解
978-7-5624-4303-2

59 质化方法在教育研究中的应用(第2版)
978-7-5624-4349-0

60 复杂调查设计与分析的实用方法(第2版)
978-7-5624-4290-5

61 研究设计与写作指导:定性、定量与混合研究的路径
978-7-5624-3644-7

62 做自然主义研究:方法指南
978-7-5624-4259-2

63 多层次模型分析导论(第2版)
978-7-5624-4060-4

64 评估:方法与技术(第7版)
978-7-5624-3994-3

65 焦点团体:应用研究实践指南(第3版)
978-7-5624-3990-5

66 质的研究的设计:一种互动的取向(第2版)
978-7-5624-3971-4

67 组织诊断:方法、模型和过程(第3版)
978-7-5624-3055-1

68 民族志:步步深入(第2版)
978-7-5624-3996-7

69 分组比较的统计分析(第2版)
978-7-5624-3942-4

70 抽样调查设计导论(第2版)
978-7-5624-3943-1

71 定性研究(第3卷):经验资料收集与分析的方法(2版)
978-7-5624-3944-8

72 定性研究(第4卷):解释、评估与描述(第2版)
978-7-5624-3948-6

73 定性研究(第1卷):方法论基础(第2版)
978-7-5624-3851-9

74 定性研究(第2卷):策略与艺术(第2版)
978-7-5624-3286-9

75 社会网络分析法(第2版)
978-7-5624-2147-4

76 公共政策内容分析方法:
978-7-5624-3850-2

77 复杂性科学的方法论研究
978-7-5624-3825-0

78 社会科学研究:方法评论
978-7-5624-3689-8

79 论教育科学:基于文化哲学的批判与建构
978-7-5624-3641-6

80 科学决策方法:从社会科学研究到政策分析
7-5624-3669-0

81 电话调查方法:抽样、筛选与监控(第2版)
7-5624-3441-7

82 研究设计与社会测量导引(第6版)
978-7-5624-3295-1